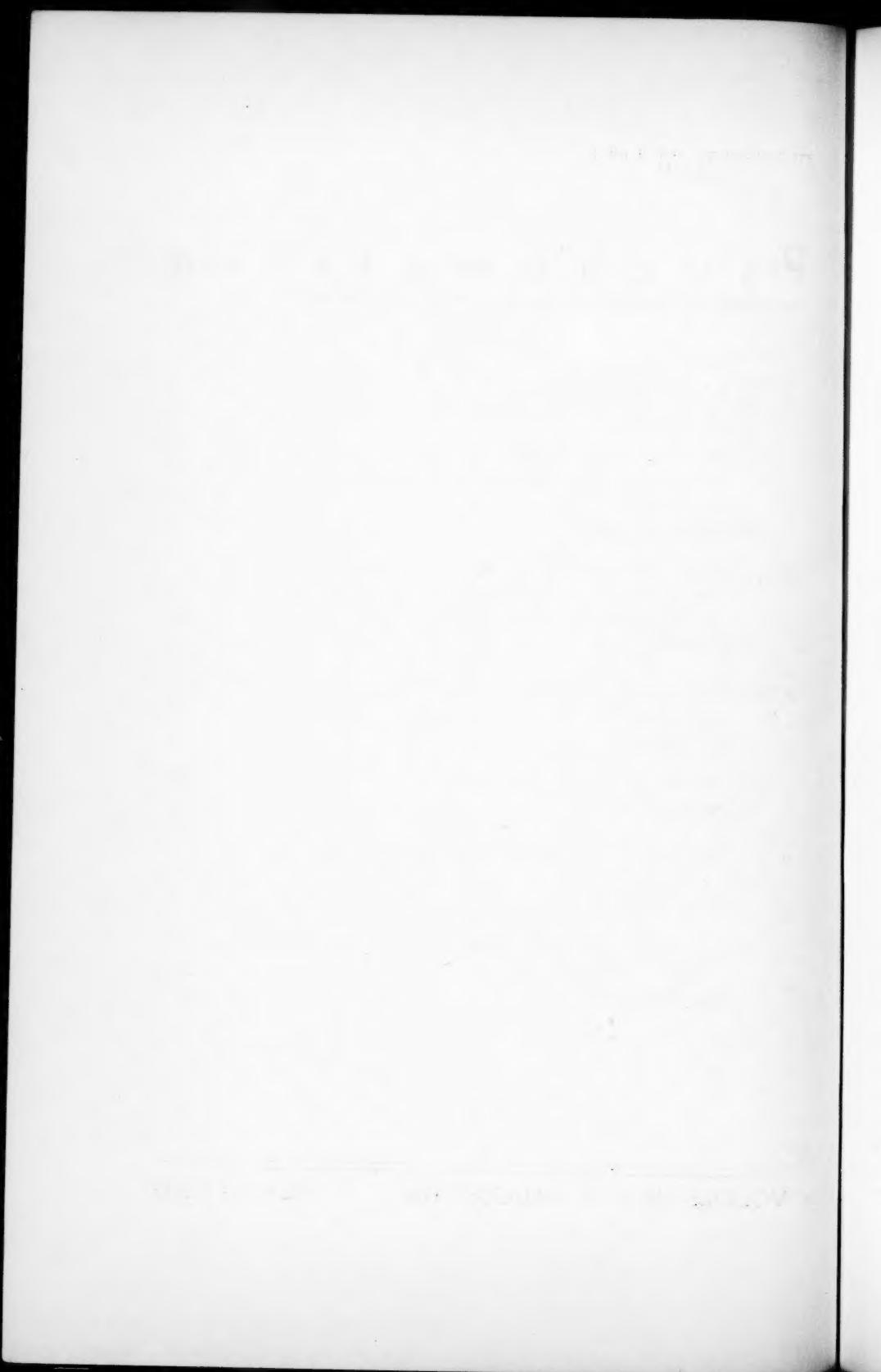


P s y c h o m e t r i k a

CONTENTS

A TECHNIQUE FOR CORRELATING MEASURABLE TRAITS WITH FREELY OBSERVED SOCIAL BEHAVIORS - - - - -	209
CHARLES C. PETERS	
THE ANALYSIS OF VARIANCE AND COVARIANCE TECHNIQUES IN RELATION TO THE CONVENTIONAL FORMULAE, ETC. - - - - -	221
MAX D. ENGELHART	
A MICRO-FILM PROJECTOR METHOD FOR PSYCHOLOGICAL TESTS - - - - -	235
L. L. THURSTONE	
THE L-METHOD - - - - -	249
HERBERT A. TOOPS	
A CRITERION FOR SIGNIFICANT COMMON FACTOR VARIANCE - - - - -	267
CLYDE H. COOMBS	
ON THE VARIATION OF THE STRUCTURE OF A SOCIAL GROUP WITH TIME - - - - -	273
N. RASHEVSKY	



A TECHNIQUE FOR CORRELATING MEASURABLE TRAITS WITH FREELY OBSERVED SOCIAL BEHAVIORS

CHARLES C. PETERS

THE PENNSYLVANIA STATE COLLEGE

In much research in social psychology it is impractical to get quantitative measure of the degree of effectiveness of certain social behaviors, yet associates can sense that effectiveness sufficiently well to detect those who manifest the behavior in very high or in very low degree. This paper develops a technique of biserial correlation from wide-spread classes to deal statistically with such situations, develops standard error formulas for it, and points to a wide range of usefulness for this type of technique.

In sociology and social psychology research situations are often encountered in which the appropriate inclusion of individuals in broad classes can be detected by their associates on the basis of long-time observation but in which finely scaled quantitative measurements can not be secured. This allocation to classes is especially feasible if the classification is into widely separated classes constituting the extreme ends of the distribution and only those individuals need to be classified who exhibit the behavior in question to such pronounced extent as to belong to one or the other of those extreme classes. If, then, we can subdivide the subjects in each of these widely separated classes dichotomously in respect to some other trait in which we are interested, or can get quantitative measurements in such trait, we can investigate the correlation between that trait and the behavior on the basis of which the aforementioned wide-spread classification was based. The former of these alternatives calls for tetrachoric correlation from wide-spread classes and the latter for biserial correlation from wide-spread classes. The former of these techniques the author has treated at considerable length elsewhere;* the latter is the subject of this paper. We shall derive a formula for biserial r for the case where only the tails of a (normal) distribution in one variable are considered, shall derive formulas for the standard error of this r , and shall illustrate its application in research. We shall develop the technique in connection with a particular research in which it is applied, which will give a clear picture of the requirements of the technique.

* Peters, C. C., and VanVoorhis, W. R. *Statistical procedures and their mathematical bases*, New York: McGraw-Hill, 1940, 375-384.

Thereafter we shall point to a number of other types of research to which the technique is also applicable. In addition to feasibility where procedures demanding finely graduated measurements would fail, the method will be shown to be highly economical because it gives much lower standard errors for the number of individuals who must be actually investigated because of taking these from the tails of the distribution rather than from the whole sample. This is very important in social research, especially when out-of-school adults are being studied, because the measurement of such individuals is an expensive process.

One of our graduate students, Mr. Joseph Krupa, was confronted with the problem of investigating the validity of such personality inventories as those of Bernreuter, of Link, and of Bell. These instruments take measurements of personality traits from testimony given by individuals about themselves. How highly do the measurements of personality traits thus obtained correlate with measurements of the actual functioning of these personalities in normal social life? It is difficult to find a technique for satisfactorily studying the validity of these instruments because it is difficult to get any dependable criterion against which to correlate the inventory scores. Ratings by outsiders based upon intimate observation of the subject's social behavior would do, but we can not ordinarily find raters who know a sufficient number of subjects well enough, and we find it particularly difficult to get a number of judges who know the *same* subjects sufficiently intimately to rate them competently. Yet it is only from the average of several judges that ratings can come to have satisfactory reliability and validity so that they may be employed as criteria.

To get an answer to this problem the following technique was applied. To 450 male college freshmen Mr. Krupa administered the Bernreuter *Personality Inventory*, the Bell *Adjustment Inventory*, and the Link *Inventory of Activities and Interests*. These subjects plus some additional freshmen (total 605) were also given ballots on which were printed paragraphs describing the characteristics and behaviors of certain hypothetical persons. These paragraphs were so constructed that one member of a pair described a person who was very high in respect to a certain trait supposedly measured by the inventories (e.g. "dominant" or "neurotic") and the other very low. The 605 freshmen were asked to write opposite each paragraph the names of one or more classmates whom they had observed to be much like the person described in the paragraph. This was done for each of the traits alleged to be measured by the three inventories. This is the so-called "guess-who" type of test.

The nominations were then tabulated under each of the traits separately. Those persons who received three or more nominations

as being like the "neurotic" character in the paragraph were taken as constituting the lower tail in the distribution in respect to neuroticism-stability while those who received three or more nominations as like the "stable" character were taken as constituting the upper tail. If the same individual received nominations for both of two contrasted classes, he was classified according to the net vote. In the case of the *B1-N* scoring of the Bernreuter Inventory, for example, the lay-out was as shown in Figure 1.

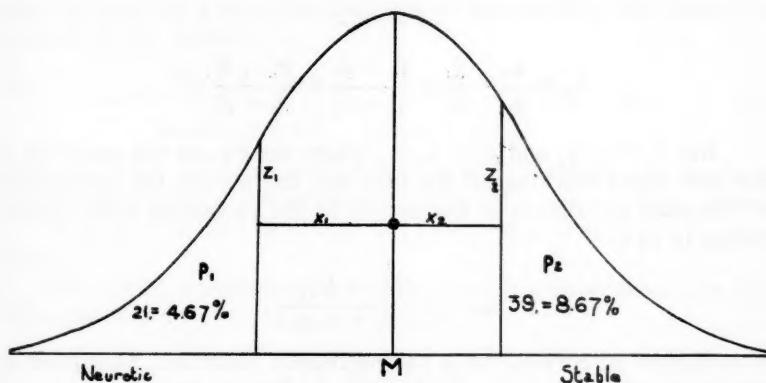


FIGURE 1

Nominations were tallied only for the 450 freshmen to whom the personality inventories had been administered. If the entire 450 cases had been included in one tail or the other, so that the whole population would have been divided into the more stable and the less stable, the ordinary bi-serial r technique could be employed to compute a validity coefficient between measurements of the trait by self-testimony and detection of the trait by observation by companions. The regular form may not, however, be used when the middle of the distribution is omitted. But a suitable formula is very easily derived, as follows.

Let the inventory scores be laid off on the y -axis and the neuroticism-stability variate as observed by companions be conceived as a *normal distribution* of unit area and unit standard deviation on the x -axis. Let \bar{x}_2 be the distance from the mean of this whole distribution to the mean of the upper tail of this distribution represented by the 39 cases nominated as highly stable. Similarly, let \bar{x}_1 be the distance from the mean of the whole distribution to the mean of the lower tail. Let \bar{y}_2 and \bar{y}_1 be corresponding statistics for the inventory scores. Then if b_{yz} is the regression coefficient,

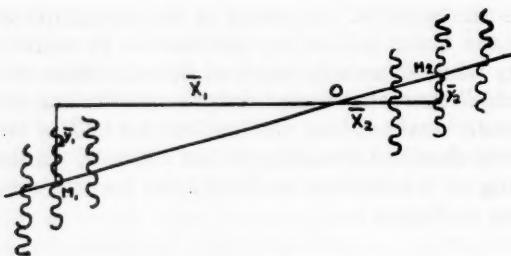


FIGURE 2

$$b_{yz} = \frac{\bar{y}_2}{\bar{x}_2} = \frac{\bar{y}_1}{\bar{x}_1} = \frac{\bar{y}_2 + \bar{y}_1}{\bar{x}_2 + \bar{x}_1} = \frac{M_2 - M_1}{\bar{x}_2 + \bar{x}_1}. \quad (1)$$

But $\bar{x}_2 = z_2/p_2$ and $\bar{x}_1 = z_1/p_1$, where the z 's are the ordinates of the unit curve marking off the tails and the p 's are the percentages of the total population of the sample in the respective tails. Substituting in (1),

$$b_{yz} = \frac{(M_2 - M_1)}{(z_2/p_2 + z_1/p_1)}. \quad (2)$$

A coefficient of correlation is the regression coefficient multiplied by the sigma ratio. That is, $r_{yz} = b_{yz}(\sigma_x/\sigma_y)$. But, by hypothesis, $\sigma_x = 1$. Whence

$$r_{b_{yz}} = \frac{(M_2 - M_1)}{\sigma_y(z_2/p_2 + z_1/p_1)} = \frac{(M_2 - M_1)p_2p_1}{(p_1z_2 + p_2z_1)\sigma_y}. \quad (3)$$

The σ_y is the standard deviation of the whole unmutilated sample of inventory scores, not that of the remaining tails. It must either be computed from the whole sample or inferred from the remaining tails. We shall return to this later.

In the case of our illustration,

M_2 is the mean $B1-N$ score of the stable group, which was 68.19;

M_1 is the mean score of the neurotic group, which was 31.0;

p_2 is the proportion of the whole population constituting the stable group, $39/450 = .0867$;

p_1 is the proportion of the whole sample in the neurotic group, .0467;

z_2 and z_1 are the ordinates of a normal curve of unit area and unit standard deviation marking off the two tails, and found from tables to be .1578 and .0977, respectively;

σ_y is the standard deviation of the whole sample of 450 *B1-N* scores, which was 30.059.

Substitution of these values in the formula gives $r = +0.316$.

We need now to know whether this is a reliable r or whether it might have arisen by chance when the true correlation is zero. For this purpose we need a formula for its standard error. We shall derive a standard error formula for the general case, then reduce it to the special case where the true correlation is assumed to be zero. We shall work in terms of a sampling universe of estimates of the true $r(\hat{r})$ inferred by the formula

$$\hat{r}_{bis} = \frac{(M_2 - M_1)p_2 p_1}{\tilde{\sigma}_y(p_1 z_2 + p_2 z_1)} = \frac{p_2 p_1}{\tilde{\sigma}_y(p_1 z_2 + p_2 z_1)} (M_2 - M_1), \quad (4)$$

where the tilde over the symbol means that it stands for the true (population) value. The coefficient of $(M_2 - M_1)$ is a constant, so that $\sigma_{\hat{r}} = k \sigma_{(M_2 - M_1)}$. Our task is to find $\sigma_{(M_2 - M_1)}$ in terms of population parameters.

The classes constituting the two tails of the distribution are uncorrelated, whence

$$\sigma_{(M_2 - M_1)}^2 = \frac{\tilde{\sigma}_2^2}{n_2} + \frac{\tilde{\sigma}_1^2}{n_1}. \quad (5)$$

For any column, a , in the distribution

$$\tilde{\sigma}_{c_a}^2 = \frac{\sum y_a^2}{n_a} - d^2,$$

where the y 's are deviations from the population mean of the class, the summation runs from $n = 1$ to $n = n_a$, and the d is the distance from this true column mean to the true mean of the whole distribution. If the regression is rectilinear,

$d = \tilde{b} x_a$ and $d^2 = \tilde{b}^2 x_a^2$. Substituting in (5),

$$\tilde{\sigma}_{c_a}^2 = \frac{\sum y_a^2}{n_a} - \tilde{b}^2 x_a^2; \quad \sum y_a^2 = n_a \tilde{\sigma}_{c_a}^2 + n_a \tilde{b}^2 x_a^2. \quad (6)$$

Summing for the whole of the upper tail (designated 2), using f to designate the frequency of a column, and assuming homoscedasticity,

$$\sum y_2^2 = n_2 \tilde{\sigma}_{c_2}^2 + \tilde{b}^2 \sum f x_2^2. \quad (7)$$

Subtracting from both sides of the equation a quantity that will leave an expression equal to $n_2 \tilde{\sigma}_2^2$ on the left,

$$\sum y_2^2 - n_2 M_2^2 = n_2 \tilde{\sigma}_c^2 + \tilde{b}^2 \sum f x_2^2 - n_2 M_2^2. \quad (8)$$

We need to make certain substitutions in (8) for purposes of simplification. $\sum f x_2^2$ can be replaced by an integral, which must be integrated by parts:

$$\sum f x_2^2 = N \int_{z_2}^{\infty} zx^2 dx = N \left[z_2 x_2 + \int_{z_2}^{\infty} z dx \right] = N(z_2 x_2 + p_2), \quad (9)$$

where the unbarred x is the distance from the mean to the point of truncation. If ρ stands for the true correlation,

$$\tilde{\sigma}_c^2 = \tilde{\sigma}_y^2(1 - \rho^2) \quad (\text{Standard error of estimate}) \quad (10)$$

$$\tilde{b} = \rho (\tilde{\sigma}_y / \tilde{\sigma}_x) = \rho \tilde{\sigma}_y \quad (\text{Since } \tilde{\sigma}_x = 1)$$

$$M_2^2 = \tilde{b}^2 \bar{x}^2 = \tilde{b}^2 (z_2^2 / p_2^2) = \rho^2 \tilde{\sigma}_y^2 (z_2^2 / p_2^2) \quad (11)$$

$$n_2 = p_2 N.$$

Making all of these substitutions in (8),

$$n_2 \tilde{\sigma}_2^2 = n_2 \tilde{\sigma}_y^2(1 - \rho^2) + N \rho^2 \tilde{\sigma}_y^2 (z_2 x_2 + p_2) - n_2 \rho^2 \tilde{\sigma}_y^2 (z_2^2 / p_2^2). \quad (12)$$

Dividing (8) through by n_2^2 and noting that $n_2 = p_2 N$,

$$\frac{\tilde{\sigma}_2^2}{n_2} = \frac{1}{N p_2} \tilde{\sigma}_y^2 (1 - \rho^2) + \frac{1}{N p_2^2} \rho^2 \tilde{\sigma}_y^2 (z_2 x_2 + p_2) - \frac{\rho^2}{N p_2^3} \tilde{\sigma}_y^2 z_2^2. \quad (13)$$

Similarly, if the correctly signed value of x_1 is used,

$$\frac{\tilde{\sigma}_1^2}{n_1} = \frac{1}{N p_1} \tilde{\sigma}_y^2 (1 - \rho^2) + \frac{1}{N p_1^2} \rho^2 \tilde{\sigma}_y^2 (p_1 - z_1 x_1) - \frac{\rho^2}{N p_1^3} \tilde{\sigma}_y^2 z_1^2. \quad (14)$$

Substituting (13) and (14) in (5),

$$\begin{aligned} \sigma^2_{(M_2 - M_1)} &= \frac{\tilde{\sigma}_y^2}{N} \left[\frac{p_1 + p_2}{p_1 p_2} (1 - \rho^2) + \frac{\rho^2}{p_1^2} (p_1 - z_1 x_1) \right. \\ &\quad \left. + \frac{\rho^2}{p_2^2} (z_2 x_2 + p_2) - \frac{\rho^2 z_2^2}{p_2^3} - \frac{\rho^2 z_1^2}{p_1^3} \right]. \end{aligned}$$

This simplifies algebraically to the following:

$$\sigma^2_{(M_2 - M_1)} = \frac{\tilde{\sigma}_y^2}{N p_1 p_2} \left[(p_1 + p_2) - \rho^2 \left(\frac{p_2 z_1^2}{p_1^2} + \frac{p_1 z_2^2}{p_2^2} + \frac{p_2 x_1 z_1}{p_1} - \frac{p_1 x_2 z_2}{p_2} \right) \right].$$

Multiplying this by the constant k in formula (4) and taking the square root,

$$\sigma_r^* = \frac{\sqrt{p_1 p_2}}{(p_1 z_2 + p_2 z_1) \sqrt{N}} \sqrt{(p_1 + p_2) - p^2 \left(\frac{p_2 z_1^2}{p_1^2} + \frac{p_1 z_2^2}{p_2^2} + \frac{p_2 x_1 z_1}{p_1} - \frac{p_1 x_2 z_2}{p_2} \right)}, \quad (15)$$

where x is the distance from the mean to the point of truncation, correctly signed. If the true correlation is zero, this simplifies to

$$\sigma_r^* = \frac{\sqrt{p_1 p_2}}{(p_1 z_2 + p_2 z_1) \sqrt{N}} \sqrt{(p_1 + p_2)}. \quad (16)$$

When the whole distribution is present this reduces to the well-known formula for the standard error of biserial r as derived by Soper, when zero is substituted for r .

Since in a universe of samples ($M_2 - M_1$) is somewhat correlated with σ_y , the standard deviation of the biserial r 's from such a supply of samples would be a little less than that for a corresponding supply of \hat{r} 's. But the difference would not be great, so that formulas (15) and (16) are good enough for practical use.

In testing the significance of our obtained r , there are two questions of interest. The first is whether our obtained r might have arisen by chance fluctuation in sampling when the true correlation is zero. For this test we need formula (16). For the data of our illustration it yields a standard error of .069. Since the r is 4.6 times this standard error, as large an r as .316 could not reasonably be attributed to chance fluctuation; so it must be admitted that there is a positive correlation between the scores on the inventory and the observation of behavior by companions.

The second question is: Within what limits may the true r between these factors be expected to lie? The answer to that question requires the standard error of r around the true r without the assumption that the true r is zero. The formula for this is (15). In applying the formula it must be remembered that the x_1 and the x_2 are the distances from the mean of a unit distribution to the points of truncation by the ordinates which mark off the respective tails, not the distances to the means of the classes in the tails. Their values are readily accessible in published tables. In formula (15) they are to be taken with their proper algebraic signs; it must be remembered that x_1 is very likely to have the negative sign in its own right, so that in actual operation the sign preceding the term containing it is likely to be

minus. For formula (15) the standard error for the problem of our illustration is found to be .066.

While it is true that the distribution of sample r 's around the true r is not strictly normal at any point, it is not far from normal when the true r is low and the population large.*

So here we may employ the normal curve function to interpret the fiducial limits of our r . Employing this type of interpretation, the chances are even that the true r lies somewhere between +.272 and +.360; and the odds are two to one that it is not lower than +.250 nor higher than +.382.

The procedure we described assumed that neuroticism-stability makes a continuous normal distribution and that there is a tendency for more companions to notice a trait to the extent to which that trait is present in great degree. So the upper tail would begin with persons who have a fair degree of stability and continue (with larger numbers of mentions) to persons who have the trait of stability to a very marked degree. An analogous thing, though oriented in the opposite direction, would be true for the lower tail. We have also assumed that the pupils know the whole 450 men well enough so that anyone who had the trait to a marked degree would have been noticed as possessing it by at least three of his companions. To the extent to which there were persons in the population not well enough known for that to happen, the r would be too low. So our obtained r is to be taken as the minimum one. In order to get what would probably be a maximum r , we assumed that 100 of the 450 men might be too little known to elicit three or more nominations even though they possessed the trait in considerable degree. We then recomputed the r with N being considered 350 instead of 450. The resultant maximum r to correspond with our minimum of .316 was .335. From this it appears that there can be considerable discrepancy in regard to the population utilized without affecting the r very violently.

The other scorings of the Bernreuter inventory, and the several scales of the Bell and the Link inventories, were treated in the same manner. The following table gives the "minimum" and the "maximum" r 's and t (the ratio of the r to the standard error if the true r were zero) for the "minimum" r .

These validity correlations are not very high. And yet they are not too bad. With two exceptions, they are high enough to permit the inventories to detect, with considerable assurance, those individuals who are outstandingly high or outstandingly low in the traits

* We are investigating the translation of this r into its hyperbolic arc-tangent, z' , and may find in this a more satisfactory method of interpreting the fiducial limits of the r .

*"Minimum" and "Maximum" Validity Coefficients, and *t*, for the
Several Scales of Three Personality Inventories*

Scale		Min. <i>r</i>	Max. <i>r</i>	<i>t</i>
Bernreuter				
B1-N, Neurotic tendency	- - - - -	.316	.335	4.58
B2-S, Self-sufficiency	- - - - -	.148	.166	2.35
B4-D, Dominance-submission	- - - - -	.426	.454	6.09
Bell,				
Health adjustment	- - - - -	.127	.141	1.62
Social adjustment	- - - - -	.444	.472	6.69
Emotional adjustment	- - - - -	.156	.163	1.98
Link,				
Social initiative	- - - - -	.478	.525	6.16
Self-determination	- - - - -	.009	.010	.12
Economic self-determination	- - - - -	-.078	-.082	.90
Link, over-all	- - - - -	.503	.534	7.08

measured—provided these persons have marked the scale with average honesty. But the correlations are not high enough to permit close discrimination within the range. Furthermore, coefficients of correlation between the inventory scores and other functions, and the spread of means in group comparisons, must be expected to be far smaller than if the personality traits were measured with instruments of perfect validity.

In Mr. Krupa's problem the standard deviations of the scores on the *Y*-axis for the whole unmutilated sample were known because the inventories had been administered to the whole sample. But that will not usually be the case; only the subjects in the tails will be measured in the trait laid out on the *Y*-axis, since economy of measurement is one of the virtues of the method. Then the standard deviation for the whole sample must be inferred from the fragments remaining in the tails. This can be done without difficulty, as follows:

Reenter our derivation above at (7) but, since we need the standard deviation of the sample, follow through in terms of the statistics of the sample rather than in terms of population parameters. Substitute in (7) the value of $\sum fx_2^2$ from (9) and of σ_c^2 from (10) and of ρ from (11). Then

$$\sum y_2^2 = n_2 \sigma_y^2 - n_2 b^2 + Nb^2(z_2 x_2 + p_2).$$

Construct a similar integral for $\sum fx_1^2$ and add, giving to x_1 its proper algebraic sign:

$$\begin{aligned} \sum y_2^2 + \sum y_1^2 &= (n_2 + n_1) \sigma_y^2 - (n_2 + n_1) b^2 \\ &\quad + Nb^2(z_2 x_2 - z_1 x_1 + p_2 + p_1). \end{aligned}$$

Transpose and solve for σ_y^2 , remembering that $n_2 = p_2 N$ and $n_1 = p_1 N$:

$$\sigma_y^2 = \frac{\sum y_2^2 + \sum y_1^2}{n_2 + n_1} - \frac{b^2(z_2 x_2 - z_1 x_1)}{p_2 + p_1}. \quad (17)$$

The y 's here are deviations from the mean of the total sample. Usually this mean will not be known. But it can be inferred; if we assume rectilinearity of regression, the total mean will be merely the weighted mean from the two tails:

$$M = \frac{n_1 M_1 + n_2 M_2}{n_1 + n_2}.$$

We could handle formula (17) by taking the y deviations from this inferred mean. But for operational purposes it will be much more convenient to work in terms of score values (which are deviations from zero). We can readily put formula (17) in terms of score values. Where the capitals stand for score values and the lower case y 's for deviation values,

$$Y_1 = y_1 + M; \quad Y_1^2 = y_1^2 + 2M y_1 + M^2;$$

$$\sum Y_1^2 = \sum y_1^2 + 2M \sum y_1 + n_1 M^2.$$

Similarly,

$$\sum Y_2^2 = \sum y_2^2 + 2M \sum y_2 + n_2 M^2.$$

Adding,

$$\sum Y_1^2 + \sum Y_2^2 = \sum y_1^2 + \sum y_2^2 + (n_1 + n_2) M^2 + 2M(\sum y_1 + \sum y_2).$$

But because $\sum y_1$ and $\sum y_2$ are deviations from the mean which is determined from their weighted sum, the quantity in the parentheses at the extreme right above will sum to zero and the term will vanish. Whence, transposing,

$$\sum y_1^2 + \sum y_2^2 = \sum Y_1^2 + \sum Y_2^2 - (n_1 + n_2) M^2.$$

Substituting this value in (17) and also the value given above for M , and taking the square root, we have:

$$\sigma_y = \left[\frac{\sum Y_1^2 + \sum Y_2^2}{n_1 + n_2} - \left(\frac{n_1 M_1 + n_2 M_2}{n_1 + n_2} \right)^2 - \frac{b^2(z_2 x_2 - z_1 x_1)}{p_1 + p_2} \right]^{\frac{1}{2}}. \quad (18)$$

So in dealing with a problem in which the standard deviation of the total sample is unknown, the first step is to determine b from formula (2); then, armed with this information, infer σ_y from formula (18); and, finally, substitute this in formula (3) to find r .

The procedure developed in this article can have wide application,

especially in social psychology. We have employed it to investigate the correlation between the extent of going in for particular areas of study in college (science, mathematics, etc.) and the possession of certain functioning cultural abilities as manifested in college life, and also to investigate the correlation between the extent of participation in certain elements of college life and achievement in after-college life. It can be applied to investigate the factors related to marital adjustment, to social leadership, to civic efficiency, to vocational competency, etc. It can, in fact, be applied wherever subjects can be detected, by those who have observed their behavior, as manifesting a type of behavior in high or low degree and where measurements of other traits can be obtained which are to be correlated therewith. Where measurements of these y -axis traits can not be obtained but the subjects can be divided in respect to them into "high" and "low," the equally apt tetrachoric r from wide-spread classes, mentioned at the opening of this article, is available. It is true that both these forms of correlation assume, besides normality of distribution, sharp truncation at the tails, which assumption would never be fully met in the type of application we have been discussing; there would be some "slopping over" of the variates on both sides of the dividing line. But that is a feature of unreliability of measurement which obtains in practically all measurements of traits of human beings other than purely physical ones; in other measurements also some cases fall in the wrong categories because of unreliability of measurement.

The technique described here is, of course, also applicable to the ordinary correlation table, and will give exactly the same r 's as the Pearson product-moment method to the extent to which all the assumptions are fully met. Since it is theoretically the same r as the Pearson product moment one, it can be put to the same uses—simple and partial regression equation prediction, multiple factor analysis, etc. It is a more economical procedure than correlation procedures involving the whole population of a random sample, because it requires the measurement of fewer individuals to give the same reliability. To yield as low standard error as that obtained in the sample of 60 subjects involved in our illustration, there would have been required 210 subjects by the Pearson product-moment method, or 330 by the regular bi-serial method. Of course the reliability is not as high as if the whole of the sample were employed from which the tails were taken; and hence we have no intention to suggest that the technique replace the regular product-moment method where all the data required for that standard method are available.

THE ANALYSIS OF VARIANCE AND COVARIANCE TECHNIQUES IN RELATION TO THE CONVENTIONAL FORMULAS FOR THE STANDARD ERROR OF A DIFFERENCE*

MAX D. ENGELHART

DEPARTMENT OF EXAMINATIONS
CHICAGO CITY JUNIOR COLLEGES

In this paper it is demonstrated that the analysis of variance techniques yield results equivalent to the calculation of t by means of expressions based on the short or the long formula. It is also shown that the covariance technique gives results equivalent to those obtained by means of the formula for t which should be used with matched groups. The conditions necessary for equivalent results are such that the conventional formulas for t would normally be used rather than the variance or covariance techniques. However, a knowledge of the relationships described in this paper should contribute to one's understanding of the variance and covariance techniques.

In handling experimental data the following formulas have been employed:

$$\sigma_{\text{difference}} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2} \quad (1)$$

$$\sigma_{\text{difference}} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2 r_{12} \sigma_{M_1} \sigma_{M_2}} \quad (2)$$

$$\sigma_{\text{difference}} = \frac{\sigma_{\text{distribution of individual differences between pairs of scores}}}{\sqrt{N}} \quad (3)$$

$$\sigma_{\text{difference}} = \sqrt{(\sigma_{M_1}^2 + \sigma_{M_2}^2)(1 - r_{12}^2)}. \quad (4)$$

Formula 1 is employed where the groups used in an experiment are independent random samples. Formula 2 is used when there is correlation between the means of successive samples, or between the paired scores of the population. Formula 2 and Formula 3, for the same data, give the same value for the standard error of the difference between the means of two groups. Formula 4 is the one used where groups have been matched. The need of such a formula was recognized by Lindquist (4) and a rigorous proof was contributed by Wilks (11).

* The relationships described in this paper were brought to the attention of the author by the excellent article of Eugene Shen (8). The proofs given here and the examples are the work of the present author.

A simple proof was given by Lindquist in his paper introducing the formula. The correlation coefficient, r_{if} , refers to the correlation between the measures used in matching (the initial measures i) and the final measures f . Formula 4 is equivalent to formulas 2 and 3 if all of the correlation between the pairs of final measures is due to what the paired individuals have in common at the start of the experiment, i.e., the common elements used as the bases of matching; if this correlation is the same within each group; and if the variance of the final measures within each group is also the same. Then

$$\sigma_{M_1}^2 = \sigma_{M_2}^2 = \sigma_{M_f}^2$$

$$r_{i1} = r_{i2} = r_{if}$$

$$r_{12,i} = \frac{r_{12} - r_{i1} r_{i2}}{\sqrt{(1 - r_{i1}^2)(1 - r_{i2}^2)}} = 0$$

$$r_{12} = r_{i1} r_{i2} = r_{if}^2.$$

Hence, formula 2 becomes

$$\begin{aligned} \sigma_{\text{difference}} &= \sqrt{2 \sigma_{M_f}^2 - 2 r_{12} \sigma_{M_f}^2} \\ &= \sqrt{2 \sigma_{M_f}^2 (1 - r_{12})} \\ &= \sqrt{2 \sigma_{M_f}^2 (1 - r_{if}^2)}, \end{aligned}$$

which under the conditions referred to is equivalent to formula 4.

Shen (8) has shown that when a single determination is made of the variance within the groups and when degrees of freedom are used in place of N , or the number of observations, the formulas for t corresponding to formulas 1, 2 and 3, and 4 are as follows:

$$t = \frac{M_{x_1} - M_{x_2}}{\sqrt{\frac{\sum (X_1 - M_{x_1})^2 + \sum (X_2 - M_{x_2})^2}{N_1 + N_2 - 2} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}} \quad (5)$$

$$t = \frac{M_{x_1} - M_{x_2}}{\sqrt{\frac{\sum (X_1 - M_{x_1} - X_2 + M_{x_2})^2}{N(N-1)}}} \quad (6)$$

$$t = \frac{M_{x_1} - M_{x_2}}{\sqrt{\frac{\sum (X_1 - M_{x_1})^2 + \sum (X_2 - M_{x_2})^2}{N_1 + N_2 - 3} (1 - r_{if}^2) \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}} \quad (7)$$

In formula 6, N refers to the number of pairs of observations and, hence, is equal to N_1 or N_2 . In formula 7, y refers to the initial measures used as the basis of matching. Formula 6 corresponds to formulas 2 and 3. The expression $(X_1 - M_{x_1} - X_2 + M_{x_2})$ represents a deviate from the mean of the individual differences between pairs of scores. This is somewhat more readily seen when the expression is written $[(X_1 - X_2) - (M_{x_1} - M_{x_2})]$. The same expression would also be used in calculating, for formula 3, the standard deviation of the distribution of individual differences between pairs of scores.

When using the simplest of the analysis of variance techniques, the variance ratio F is computed by means of the following formula:

$$F = \frac{\frac{n \sum (M_p - M)^2}{r-1}}{\frac{\sum d^2}{r(n-1)}},$$

where M_p is a group or class mean, M is the general mean, d symbolizes the deviations of the individual measures in all of the groups from the means of their respective groups, r is the number of groups or classes, and n is the number of individuals in each group or class. The numerator of the fraction is an estimate of the population variance based on the variance of the group means or variance between groups. The denominator is an estimate of the population variance based on the variance within groups. For a more extended discussion and derivations, see Lindquist's text (5). When there are only two groups, $F = t^2$ calculated by means of formula 5. Since $r = 2$ and 2 replaces Σ in the numerator because the summation is over 2 groups,

$$F = \frac{2 n (M_p - M)^2}{\frac{\sum d^2}{2(n-1)}}$$

$$M_p - M = \frac{M_{x_1} - M_{x_2}}{2}$$

$$(M_p - M)^2 = \frac{(M_{x_1} - M_{x_2})^2}{4}.$$

Hence,

$$F = \frac{\frac{2 n (M_{x_1} - M_{x_2})^2}{4}}{\frac{\sum d^2}{2(n-1)}} = \frac{n (M_{x_1} - M_{x_2})^2}{2} \cdot \frac{2(n-1)}{\sum d^2}$$

$$F = \frac{(M_{z_1} - M_{z_2})^2}{\frac{\sum d^2}{n(n-1)}}.$$

This is equivalent to the square of t from formula 5, since $d = X_1 - M_{z_1}$ and $X_2 - M_{z_2}$ for both groups and n equals N_1 or N_2 . The relationship of formula 5 to formula 1 has been indicated.

The within groups variance previously referred to is calculated from the within groups sum of squares. If the measures in two groups have been paired, the sum of squares for between the means of pairs can be removed from the within groups sum of squares, yielding a remainder sum of squares, which, when divided by the appropriate degrees of freedom and inserted in the formula for F yields a value for F which is equal to t^2 calculated by means of formula 6. The proof of this relationship is as follows:

For two groups the within groups sum of squares equals

$$\sum (X_1 - M_{z_1})^2 + \sum (X_2 - M_{z_2})^2.$$

Since the mean of a pair of scores equals $X_1 + X_2$ and n for the pair equals 2, the sum of squares between the means of pairs, $2 n \sum (M_p - M)^2$ becomes

$$2 \sum \left[\frac{X_1 + X_2}{2} - M \right]^2$$

or

$$\frac{2 \sum (X_1 + X_2 - M_{z_1} - M_{z_2})^2}{4},$$

since

$$M = \frac{M_{z_1} + M_{z_2}}{2}.$$

The remainder sum of squares thus equals

$$\sum (X_1 - M_{z_1})^2 + \sum (X_2 - M_{z_2})^2 - \frac{\sum (X_1 + X_2 - M_{z_1} - M_{z_2})^2}{2}.$$

Expanding the first two terms, summing, recalling that $\sum X = NM$, and multiplying by 2,

$$2 \sum X_1^2 - 2 N M_{z_1}^2 + 2 \sum X_2^2 - 2 N M_{z_2}^2.$$

Expanding the numerator of the last term, summing, and changing signs,

$$-\sum X_1^2 - \sum X_2^2 - N M_{x_1}^2 - N M_{x_2}^2 - 2 \sum X_1 \sum X_2 + 2 N M_{x_1}^2 + 2 N M_{x_2} M_{x_1} + 2 N M_{x_1} M_{x_2} + 2 N M_{x_2}^2 - 2 N M_{x_1} M_{x_2}.$$

Combining the two expansions of terms and dividing by 2,

$$\frac{\sum X_1^2 - 2 \sum X_1 \sum X_2 + \sum X_2^2 - N M_{x_1}^2 - N M_{x_2}^2 + 2 N M_{x_1} M_{x_2}}{2},$$

which equals

$$\frac{\sum (X_1 - M_{x_1} - X_2 + M_{x_2})^2}{2}.$$

For proof, expand, sum, and substitute $\sum X = N M$, etc.

The remainder variance equals

$$\frac{\sum (X_1 - M_{x_1} - X_2 + M_{x_2})^2}{2(N-1)},$$

where $(N-1)$ equals the degrees of freedom.

The variance between groups was previously shown to be equal to

$$\frac{2 N (M_{x_1} - M_{x_2})^2}{4}.$$

Hence

$$F = \frac{\frac{2 N (M_{x_1} - M_{x_2})^2}{4}}{\frac{\sum (X_1 - M_{x_1} - X_2 + M_{x_2})^2}{2(N-1)}} = \frac{(M_{x_1} - M_{x_2})^2}{\frac{\sum (X_1 - M_{x_1} - X_2 + M_{x_2})^2}{N(N-1)}},$$

which is equal to t^2 as calculated by means of formula 6, the formula corresponding to formulas 2 and 3.

The covariance techniques can be used to allow for initial inequality of groups by use of the regression of final on initial scores. One could compute the individual adjusted final scores of all individuals and apply the ordinary variance techniques; i.e., each final X score would be corrected by an amount equal to by , where b is the regression of X on Y and y represents the deviation of the individual from

the mean of the initial measures. If an individual deviates above the initial mean, correction is made by subtracting by from his final score. If he deviates below the initial mean, by is added to allow for the initial handicap. It is not necessary to do this to the individual score. If he deviates below the initial mean, by is added to allow for the initial handicap. It is not necessary to do this to the individual scores since the adjusted sums of squares can be computed and used in obtaining the adjusted between groups and within groups variances used in calculating F .

The adjusted sum of squares for total is

$$\Sigma[(X - by) - M_x]^2 = \Sigma(x - by)^2 = \Sigma x^2 - \frac{(\Sigma x y)^2}{\Sigma y^2},$$

where X is any final score, M_x is the mean of all final scores, and x and y are deviations from their respective general means.

The adjusted sum of squares for within groups is

$$\Sigma[(x - \bar{x}) - b(y - \bar{y})]^2 = \Sigma(x - \bar{x})^2 - \frac{[\Sigma(x - \bar{x})(y - \bar{y})]^2}{\Sigma(y - \bar{y})^2}.$$

The left-hand member of the above identity contains, within the bracket the deviation of any final score, x , from the final group mean, \bar{x} , minus the regression coefficient times the deviation of the corresponding initial score y from the mean of the initial scores of the group. (The group means \bar{x} and \bar{y} are expressed as deviates from the gross score general means M_x and M_y .) The summation is over all groups. The right-hand member is analogous with the final expression given for the adjusted sum of squares for total, but the b 's are different. The former is calculated from the total series of paired scores, while the latter is the average b of the several groups, a within groups regression coefficient. The adjusted sum of squares for between groups may be obtained by subtracting the adjusted sum of squares for within groups from the adjusted sums of squares for total. F is equal to the ratio between the adjusted variance between groups and the adjusted within groups variance.*

If two groups are matched on the basis of initial measures, one would not use the covariance procedure. However, if one does use the covariance procedure, the resulting F is equal to t^2 calculated from formula 7, the formula which is related to formula 4.

From the preceding discussion and for two groups

* For a more extended discussion of the covariance techniques see Lindquist's text (5). In Lindquist's treatment X represents the initial and Y the final scores. The symbols are reversed here for consistency within this article.

$$F = \frac{\frac{\left(\sum x^2 - \frac{(\sum \bar{x} y)^2}{\sum y^2} \right) - \left(\sum (x - \bar{x})^2 - \frac{[\sum (x - \bar{x})(y - \bar{y})]^2}{\sum (y - \bar{y})^2} \right)}{N_1 + N_2 - 3}}{\frac{1}{\sum (x - \bar{x})^2 - \frac{[\sum (x - \bar{x})(y - \bar{y})]^2}{\sum (y - \bar{y})^2}}}.$$

x and y are deviation measures from the general means M_x and M_y . x and \bar{y} are the group means, M_{x_1} and M_{x_2} and M_{y_1} and M_{y_2} , expressed for both groups as deviates from the general means M_x and M_y . The summations indicated above are over both groups. Now, if the groups are identically matched on the basis of the Y scores,

$$\frac{\sum y_1}{N_1} = \frac{\sum y_2}{N_2} = 0 \quad \text{and} \quad \bar{y}_1 = \bar{y}_2 = 0.$$

Hence, $\sum (y - \bar{y})^2 = \sum y^2$ and $\sum (y - \bar{y}) = \sum y$.

For two groups of equal size, \bar{x} for one group is equal, but opposite in sign, to \bar{x} for the other group, since the x deviation measures are from the general mean M_x , which is itself the average of the raw score group means M_{x_1} and M_{x_2} . Hence, $\sum (x - \bar{x})(y - \bar{y})$ equals $\sum [x - (+\bar{x})]y$ for one group and $\sum [x - (-\bar{x})]y$ for the other group. It follows that $\sum (x - \bar{x})(y - \bar{y})$ equals $\sum x y + \bar{x} \sum y$ for one group and $\sum x y - \bar{x} \sum y$ for the other group. Since $\sum y$ is the same for both groups $\sum (x - \bar{x})(y - \bar{y})$ is simply $\sum x y$. Then

$$F = \frac{\sum x^2 - \sum (x - \bar{x})^2}{\sum (x - \bar{x})^2 - \frac{(\sum x y)^2}{\sum y^2}};$$

$$\frac{1}{N_1 + N_2 - 3}$$

$$\frac{(\sum x y)^2}{\sum y^2} = \frac{(\sum x y)^2}{\sum y^2} \cdot \frac{\sum (x - \bar{x})^2}{\sum (x - \bar{x})^2} = r_{xy}^2 \sum (x - \bar{x})^2.$$

r_{xy} is an average within groups correlation. The usual formula is

$$\frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}},$$

and the summation is over both groups. It was previously indicated that $\sum (x - \bar{x})(y - \bar{y}) = \sum x y$ and that $\sum (y - \bar{y})^2 = \sum y^2$. Then

$$F = \frac{\frac{\sum x^2 - \sum (x - \bar{x})^2}{\sum (x - \bar{x})^2 - r_{xy}^2 (x - \bar{x})^2}}{\frac{N_1 + N_2 - 3}{N_1 + N_2 - 3}} = \frac{\sum x^2 - \sum (x - \bar{x})^2}{\sum (x - \bar{x})^2 (1 - r_{xy}^2)}.$$

By definition,

$$\bar{x}_1 = M_{z_1} - M_z$$

or

$$M_z = M_{z_1} - \bar{x}_1$$

and

$$x_1 = (X_1 - M_z)$$

or

$$x_1 = (X_1 - M_{z_1} + \bar{x}_1).$$

Similarly,

$$x_2 = (X_2 - M_{z_2} + \bar{x}_2).$$

Summing over both groups

$$\sum x^2 = \sum (X_1 - M_{z_1} + \bar{x}_1)^2 + \sum (X_2 - M_{z_2} + \bar{x}_2)^2.$$

Since

$$x_1 - \bar{x}_1 = (X_1 - M_z) - (M_{z_1} - M_z) = X_1 - M_{z_1},$$

when the summation is over both groups

$$\sum (x - \bar{x})^2 = \sum (X_1 - M_{z_1})^2 + \sum (X_2 - M_{z_2})^2.$$

Since

$$(X_1 - M_{z_1} + \bar{x}_1)^2$$

$$= X_1^2 - 2 X_1 M_{z_1} + 2 X_1 \bar{x}_1 - 2 M_{z_1} \bar{x}_1 + M_{z_1}^2 + \bar{x}_1^2$$

and

$$(X_1 - M_{z_1})^2 = X_1^2 - 2 X_1 M_{z_1} + M_{z_1}^2,$$

the difference between the right-hand members of the last two identities is equal to

$$2 X_1 \bar{x}_1 - 2 M_{z_1} \bar{x}_1 + \bar{x}_1^2,$$

which, when summed for the group, is equal to

$$2 \bar{x}_1 \sum X_1 - 2 N_1 M_{z_1} \bar{x}_1 + N_1 \bar{x}_1^2.$$

Since

$$\sum X_1 = N_1 M_z,$$

the foregoing expression equals

$$2 N_1 M_{z_1} \bar{x}_1 - 2 N_1 M_z \bar{x}_1 + N_1 \bar{x}_1^2,$$

which reduces to $N_1 \bar{x}_1^2$. Hence the difference between the expressions previously given for $\sum x^2$ and $\sum (x - \bar{x})^2$ reduces for both groups to $N_1 \bar{x}_1^2 + N_2 \bar{x}_2^2$, which is equal to

$$N_1 (M_{x_1} - M_x)^2 + N_2 (M_{x_2} - M_x)^2.$$

Since

$$M_{x_1} - M_x = \frac{(M_{x_1} - M_{x_2})}{2}$$

and

$$M_{x_2} - M_x = \frac{-(M_{x_1} - M_{x_2})}{2}$$

on the assumption that M_{x_2} is smaller than M_{x_1} , the numerator F becomes

$$\frac{(N_1 + N_2)(M_{x_1} - M_{x_2})^2}{4},$$

and, since $\sum (x - \bar{x})^2$ equals

$$\sum (x_1 - M_{x_1})^2 + \sum (x_2 - M_{x_2})^2$$

for both groups and $\frac{N_1 + N_2}{4}$, when inverted, equals $\frac{1}{N_1} + \frac{1}{N_2}$,

$$F = \frac{(M_{x_1} - M_{x_2})^2}{[\sum (X_1 - M_{x_1})^2 + \sum (X_2 - M_{x_2})^2] \left(\frac{1}{N_1} + \frac{1}{N_2} \right) (1 - r_{xy}^2) \left(\frac{1}{N_1} + \frac{1}{N_2} \right)},$$

which is equal to t^2 as calculated by means of formula 7, the formula related to formula 4.

The relationships proved in the preceding paragraphs are illustrated by computations based on fictitious data given in Table 1. It will be seen that the groups are identically matched on the basis of the initial Y scores.

Analysis of variance* of final scores corresponding to the use of formula 5:

Sum of squares for total

$$\begin{aligned} \sum X^2 - G T \cdot G M &= 16^2 + 14^2 + 12^2 + \dots + 4^2 + 1^2 - 510 \cdot 8.5 \\ &= 5186 - 4335 = 851. \end{aligned}$$

* Derivations of the gross-score formulas are given in Lindquist's text.

TABLE I. Fictitious Experimental Data

	Initial		Final		Totals of Pairs of Final Scores	Means of Pairs of Final Scores
	Scores Y_1	Y_2	Scores X_1	X_2		
1	9	9	16	11	27	13.5
2	8	8	14	16	30	15
3	8	8	12	9	21	10.5
4	7	7	15	13	28	14
5	7	7	14	12	26	13
6	7	7	14	10	24	12
7	6	6	12	13	25	12.5
8	6	6	13	11	24	12
9	6	6	13	12	25	12.5
10	6	6	10	9	19	9.5
11	6	6	8	9	17	8.5
12	5	5	9	7	16	8
13	5	5	9	6	15	7.5
14	5	5	11	9	20	10
15	5	5	10	11	21	10.5
16	5	5	8	7	15	7.5
17	5	5	6	7	13	6.5
18	5	5	11	12	23	11.5
19	4	4	9	8	17	8.5
20	4	4	7	8	15	7.5
21	4	4	6	2	8	4
22	4	4	5	4	9	4.5
23	4	4	7	8	15	7.5
24	4	4	7	5	12	6
25	3	3	4	2	6	3
26	3	3	6	7	13	6.5
27	3	3	5	4	9	4.5
28	3	3	4	3	7	3.5
29	2	2	3	4	7	3.5
30	1	1	2	1	3	1.5
Total	150	150	270	240	510	255
Mean	5	5	9	8	17	8.5

Sum of squares for between groups

$$\frac{T_{z_1}^2 + T_{z_2}^2}{30} - G T \cdot G M = \frac{270^2 + 240^2}{30} - 4335 = 15.$$

Sums of squares for within groups

$$851 - 15 = 836.$$

When the between groups and within groups sums of squares are divided by 1 and 58, their respective degrees of freedom, the corre-

sponding variances are 15 and 14.41 and $F = 1.04$. Using formula 5,

$$t = \frac{9 - 8}{\sqrt{\frac{428 + 408}{30 + 30 - 2} \left(\frac{1}{30} + \frac{1}{30} \right)}} = 1.02;$$

$$t^2 = 1.04.$$

Analysis of variance of final scores with removal of the sum for between the means of pairs from the within groups sum of squares:

Sums of squares for between the means of pairs

$$\frac{T_1^2 + T_2^2 + \dots + T_{30}^2}{2} - G T \cdot G M$$

$$= \frac{27^2 + 30^2 + \dots + 7^2 + 3^2}{2} - 4335 = 789.$$

The previously obtained within-groups sum of squares, 836, minus 789 yields a remainder sum of squares of 47. Dividing this sum of squares by 29 degrees of freedom, the remainder or error variance is 1.62.

$$F = \frac{15}{1.62} = 9.25.$$

Using formula 6

$$t = \frac{9 - 8}{\sqrt{\frac{94}{30 \cdot 29}}} = 3.04 + .$$

$$t^2 = 9.25.$$

Covariance analysis:

The adjusted total sum of squares

$$851 - \frac{(348)^2}{196} = 233.$$

The adjusted sum of squares for within groups

$$836 - \frac{(348)^2}{196} = 218.$$

Adjusted sum of squares for between groups

$$233 - 218 = 15.$$

Since the degrees of freedom are 1 and 57, the adjusted variance between groups is 15 and the adjusted variance within groups is 3.82.

$$F = \frac{15}{3.82} = 3.92.$$

Using formula 7 in the calculation of t ,

$$t = \frac{9 - 8}{\sqrt{\frac{428 + 408}{30 + 30 - 3} (1 - .86^2) \left(\frac{1}{30} + \frac{1}{30}\right)}};$$

$$t = 1.98;$$

$$t^2 = 3.92.$$

The values of t calculated on the basis of formulas 6 and 7 are not the same. The difference is due to the fact that the correlation between the final scores is .89, while the square of the correlation between the initial and final scores is .74. Hence, in this hypothetical case, the groups apparently were not matched on the basis of all common factors. Had the within-groups correlation coefficient r_{xy} equaled the square root of the correlation between the final measures, the covariance technique, the t calculated by formula 7, the variance technique involving the removal of the variance between the means of pairs, and the t calculated from formula 6 would have given identical results, i.e., the F 's would have been equal, the t 's would have been equal, and the squares of the t 's would have equaled the F 's.

In closing, it should be mentioned that multiple regression coefficients can be used in making covariance adjustments for initial inequality of groups. Possibly in the future we may use regression coefficients derived from factor analysis and thus allow for initial inequality in all relevant primary abilities.

BIBLIOGRAPHY

1. Deemer, Walter L. A numerical example illustrating the generalized formula for testing significance of experimental treatments. *Harvard educ. Review*, 1940, 10, 75-81.
2. Engelhart, Max D. Classroom experimentation. *Review educ. Research*, 1939, 9, 555-563.
3. Fisher, R. A. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1938. 7th ed.
4. Lindquist, E. F. The significance of a difference between matched groups. *J. educ. Psychol.*, 1931, 22, 197-204.
5. Lindquist, E. F. Statistical analysis in educational research. Boston: Houghton Mifflin Company, 1940.
6. Rider, P. R. An introduction to modern statistical methods. New York: John Wiley and Sons, 1939.
7. Shen, Eugene. A generalized formula for testing the significance of experi-

mental treatments. *Harvard educ. Review*, 1940, 10, 70-74.

- 8. Shen, Eugene. Experimental design and statistical treatment in educational research, *J. exper. Educ.*, 1940, 8, 346-353.
- 9. Snedecor, G. W. Statistical methods. Ames, Iowa: Collegiate Press, 1938.
- 10. Walker, Helen M. Degrees of freedom. *J. educ. Psychol.*, 1940, 31, 253-269.
- 11. Wilks, Samuel S. The standard error of the means of matched samples. *J. educ. Psychol.*, 1931, 22, 205-208.

A MICRO-FILM PROJECTOR METHOD FOR PSYCHOLOGICAL TESTS

L. L. THURSTONE
UNIVERSITY OF CHICAGO

Adaptations of the micro-film projection method to the administration of both individual and group psychological tests are described, together with newly designed auxiliary equipment. The author cites examples of the application of this technique, which can be adapted to a variety of testing situations and which seems to be of unusual promise because of the attention value of visual projection.

A projector test method

In this paper we shall describe a method of conducting psychological examinations which has been in use in our laboratory for the last two years and which seems to have considerable promise for clinical use. The method is efficient as regards the time of both subject and examiner. It is economical in cost because a variety of psychological test material can be adapted easily to micro-film projection; it can be adapted for different psychological test purposes, as we hope to show with a number of examples; and the method can be used effectively for individual tests and for some types of group tests.

Since the method was designed for experiments with individual subjects in the laboratory, the new method will be described first with some of the individual tests for which it has already been used.

In one test of a series of perceptual tests that we are now investigating, the subject is presented single words that have been mutilated by the elimination of a part of each letter. These are presented to the subject, one word at a time, and the subject is asked to pronounce the word as quickly as he can read it. The score in the test consists of the median response time for a set of twenty mutilated words. A record is also kept of each word that the subject fails to identify in half a minute.

If this were the only test in the battery, or if they were all of the same general kind, as regards test method, the mutilated words could be drawn on cards which could be turned over by the examiner in front of the subject. The examiner could operate a stop watch at the moment of presentation and at the moment of the subject's response. A more formal arrangement would be a special exposure apparatus for the test. As we shall see, however, a variety of perceptual tests might require a number of tachistoscopic devices. These can be com-

bined into a single tachistoscopic unit in the present equipment which is adaptable to a variety of tests.

The mutilated words to be used in the test are mounted on cards of any convenient size, one word on a card, such as 3"x5", or 4"x6". The instruction material is typewritten on one or more cards. The cards are arranged in the proper order, and they are sent to a micro-film laboratory such as are maintained by a number of libraries. The photographic record consists of a negative 35 mm. film on which each card has been photographed, one card for each single frame on the film. If it is desired to present the material as white on black to the subject, one uses the negative film. If one prefers to have black on white, a positive film copy is obtained. These film strips are quite inexpensive. The cost is not more than two cents per frame, and it is less for copies of the first negative film.

The tests are given in a room which can be either completely darkened or partially darkened. Some tests can be given in ordinary room light. We have used an S.V.E. projector, as shown at *A* in *Figure 1*. This projector is Model AAA, 300 watts. The screen, not shown in the photograph, is mounted on a table to the right of the apparatus in the photograph of *Figure 1*. The subject sits near the table at *R*, and the examiner sits to his right at *T*. In this position the examiner has ready access to all of the controls of the apparatus. The subject gives his responses to these tests in one of two ways. He is instructed to press the right-hand key or the left-hand key at *N* to indicate his response, or he is instructed to give his response orally in the voice key at *M*. In the test of reading mutilated words he is asked to respond by pronouncing the word into the voice key as quickly as he can.

The procedure in giving this test is briefly as follows: The room is darkened sufficiently for projection purposes, and the light of the projector is turned off. A pilot light which is mounted at the top of the portable set *H* gives sufficient light for the examiner to operate the controls. He advances the film one frame. Then he gives the warning signal, "Ready," to the subject. Within one to three seconds he presses the key *Q* which turns on the light of the projector and also starts the Standard timer at *H*. The word is then visible to the subject on the screen in white or black. As soon as the subject reads the word into the microphone at *M*, the light of the projector is automatically turned off and the timer is automatically stopped. The examiner notes the time and the subject's response. The examiner turns the film ahead to the next frame, and the procedure is repeated. We have found that it is much easier to maintain an active interest and attention of the subject with the projector method than when the same material is presented either as a printed list or on cards.

Several tests are spliced on the same film so that only one strip of film needs to be handled for each sitting. In order to illustrate how the instructions for the test are arranged on film, we reproduce here the text that was typewritten on successive cards for this purpose. The following list shows the text on each of the instruction frames. The examiner advances the film, one frame at a time, and he reads the instructions slowly to the subject. All of the screen tests are presented in this way with instructions and examples on as many frames as are required.

Text projected on successive frames:

1. Mutilated Words
2. On each frame you will see a word.
Parts of the word have been erased.
See how quickly you can pronounce it.
3. (Sample of mutilated word)
4. All of the words used in this test are ordinary words.
5. The test begins on the next frame.
Pronounce each word as promptly as you can.

One of the tests that was prepared for the manual response keys is called "Circles." The psychological idea of this test is that there are individual differences in the amount of retardation of response when a difficult decision is to be made. Discrimination is slower for difficult comparisons than for easy comparisons. In order to equalize the attention value of the stimulus as far as possible, we have used a simple psychophysical discrimination, namely, to judge which of two disks on the screen is the larger in diameter. A similar test consists in deciding which of two squares has the larger number of dots. The two stimuli are arranged horizontally on the screen. The subject is instructed to press the right-hand key if the right-hand stimulus is the larger and to press the left-hand key if the left-hand stimulus is the larger. Some of the discriminations are extremely easy so that the response time approaches reaction time, while some of the stimulus pairs differ only slightly so as to require closer examination. It is an important consideration to arrange the stimuli in relation to the two response keys so that the instructions are easy and natural. We have preferred to arrange paired stimuli so as to give a natural right and left comparison with a response to the larger, the stronger, or the more desirable stimulus. With this arrangement, the subject need never spend any time recalling just what the instructions were. He falls naturally into the mode of response.

Other tests in this battery which are given with the equipment described so far are as follows: Dotted Outlines, Street Gestalt Completion Test, and Social Judgment. In Dotted Outlines the subject sees a digit in a field of spots very much as in the Stilling charts for testing color blindness which are also used in the Ishihara color-blindness test. Each frame shows a field of white spots on a dark background. The subject responds with the voice key. In Dotted Outlines the subject is asked to complete a figure by imaginary straight lines between dots so as to produce a capital letter or a digit. The Street Gestalt Completion Test is presented in a complete form with thirty pictures, one at a time, and the subject gives his response in the voice key. In the test called Social Judgments the subject is shown pairs of words, one pair on each frame, and he is asked to choose one as the socially more desirable. The pairs are so chosen that one of each pair denotes an individualistic trait while the other is more social. The subject is unaware of this pattern. The scores here are the median response time and the number of social traits chosen. A few sample pairs are: careless-inconsiderate, competent-tactful, tolerant-calm, stupid-cruel. If the object-person preference in word association has any psychological significance for the study of individual differences, it should be possible to find it in terms of a test with a single word on each frame, the subject being asked to respond with a short phrase or sentence using the word or a free association with the word. Each word in the test is chosen because it has two meanings, one of which denotes something physical, literal, objective, while the other meaning denotes something more personal, human, social. Which type of association comes most readily to the subject who is asked to respond quickly with an association or phrase using the word? A few sample words are: fire, slight, yield, concern, tear, friction, and pull.

In a test of individual differences in peripheral span, modification is introduced into the procedure. In order to cover a wider range on the screen, the test material is photographed on double frames which have the film size 1" x 1 1/2", that is well known for miniature cameras, such as the Leica and the Contax. In order to project double frames with the projector at *A*, in the photograph of *Figure 1*, the film carriage is turned through a right angle. This is the only change required in the projector. The subject is given a fixation point on the screen by means of the smaller auxiliary projector at *F*. The film in that projector is merely a dense field with a transparent spot at the center. The auxiliary projector is mounted so that it can be readily adjusted to place the fixation point at the center of the field from the main projector *A*. The film in the projector *A* has only one letter in each double frame. Immediately in front of the projector is placed

a shutter at *B*, which can be exposed for 1/25 second. This is an Ilex shutter. While this shutter was being repaired, another simple pendulum shutter was designed by Mr. Albert Hunsicker. It is shown at *C*. It is silent and easily operated by the examiner from his position at *T*. The detailed design of Hunsicker's pendulum shutter may be separately published. The light of the projector *A* is left on continuously in this test, and the short exposure of each double frame is made by means of the shutter at *B* or the pendulum shutter at *C*. When the subject is given the warning signal, he attends to the fixation point. When the exposure is made, there appears on the screen a capital letter for 1/25 second. The letter appears somewhere in the periphery but in an unknown direction from the fixation point as far as the subject is concerned. The subject pronounces the letter if he can read it. The successive double frames so projected show letters at increasing distances from the fixation point. Six different distances from the fixation point are used with four different readings at each distance.

Another variation in procedure is introduced in a test to determine gross individual differences in dark adaptation time. This test is not intended to give an accurate or complete determination of dark adaptation time. This test is only intended to give a rough index for each subject in studying individual differences in dark adaptation time. The 500-watt lamp at *E* is turned on so as to illuminate the screen. The subject is asked to look at the center of the screen. At the same time, the projector *A* is projecting a capital letter on the screen, but this letter is not visible because of the strong lamp *E*. A single switch controls the light *E* and also the timer *H*. When the light *E* is turned off, the timer *H* automatically starts. The subject looks at the screen until he can read the letter which is projected in faint illumination. When the subject reads the letter into the voice key, the timer automatically stops, and his adaptation time is read on the timer. This time is an index which can be used for comparative study of dark adaptation time in relation to other perceptual functions. Ten readings are made for each subject with a different letter for each reading. The films for this test are double-frame size, and they are mounted in 2" x 2" slides. The projector *A* can be used for single frame, double frame, or slides. In order to make the illumination low enough on the screen to obtain suitable individual differences in response time for this test, it was found necessary to prepare a filter with several dense negatives which were inserted in a round capsule that could be slipped over the end of the objective of the projector. In giving this test the shutters *B* and *C* are both left open for continuous projection of the letter.

Slides are also used in a test of color and form memory. Colored

slides can be prepared in one of two ways. Kodachrome film can be used in photographing colored material, or the material can be prepared as India ink drawings. The negative film for the drawings can then be tinted by hand with a fine brush. We have used both methods. In the present test the negative slides were colored by hand. The subject is again asked to attend to the fixation point which is given by the auxiliary projector. By means of one of the shutters, the examiner exposes for 1/25 second one of the ten colored slides. The subject is asked to recall the forms as well as the colors. His scores are the number of forms recalled and the number of colors recalled. Ten forms and four colors are used in this test. Before being given the test the subject is shown the ten forms in black and white, and he is asked to name them. He is also tested for color blindness by the Ishihara test.

The projector method for group tests

It seems likely that the projector method of conducting psychological examinations will also be found useful for group testing. We have used this method to advantage in conducting a factorial study of twenty-four new memory tests in combination with sixteen tests for other cognitive functions. In presenting the material to be memorized we found the projector method especially useful in that the exposure time could be controlled for each item. This procedure insures that the subjects spread their attention evenly over the material that was to be recalled. One of the most important features of the projector method of conducting psychological examinations is the ease with which one gets the undivided attention of the subjects to the presented material. In a quiet room with forty subjects, one can hear the click of the projector when the operator advances the film to the next frame which presents the next item to be studied. The expectancy of the successive items, presented with a uniform exposure time of a few seconds each, holds the attention of the subjects in a manner which would certainly be impossible if the same material were presented to them on a printed page. The attention value of the visual projector method can be regarded as one of its principal features.

We shall describe here a few of the memory tests as examples of the projector method. The tests will be described in further detail in connection with the experimental study of a battery of memory tests. A test of paired associates for words and numbers was presented by showing a word with its associated number on each frame. The operator advanced the film once every four seconds. When the entire list of paired associates had been presented, the subjects were given a printed list of the words with instruction to record as many of the associated numbers as they could recall. A list of statements was presented in the same manner, one statement on each frame. When

the entire film had been presented, the subjects were given a form of completion test in which some of the words were omitted. Another variant of this test consisted in presenting with the projector a list of statements on controversial issues. This was followed by a printed list of questions in which the subjects were asked whether the film showed any statement favoring specified opinions. This test was intended as a test of logical memory. A set of pictures was presented in a similar manner, one on each frame. Later, the subjects were given a printed form containing four times as many pictures, and they were asked to check those which had been shown. The test was intended to eliminate verbalization as far as possible. One of the pictures showed a chair, and the recall form showed pictures of four chairs. If the subject had merely relied on verbalization of the picture for recall, he was at a disadvantage, and the test featured to this extent the recall of visual detail, as such. Two tests were concerned with the recall of faces and names. Photographs were shown by the film, one on each frame, together with names. Later, the subjects were shown the pictures in a printed form, and they were asked to write the corresponding names. Another set of photographs was shown by means of the film, and the subjects were asked to check those photographs in a printed folder which had been shown previously by the film. This tests another form of visual memory in which verbalization is minimized. A set of limericks was presented by the film, one on each frame, with adequate time for reading each one. Later, the same limericks were shown to the subjects in a printed pamphlet, except that the last line was missing from each limerick. The subjects were asked to recall the last line of each limerick. A set of absurd sentences was also presented by means of the film. The recall was written by the subjects in a printed list of the same sentences, with certain words omitted. Psychophysical judgments were also made by the subjects with the projector. They were shown a square filled with small dots. This frame was followed by a blank dark frame which was kept for several seconds. Then followed another frame which contained a square filled with small dots. The subjects were asked to record on the printed forms whether the second frame contained more dots than the first. The experiment then continued to another psychophysical comparison pair. This test was given with the room sufficiently light so that the subjects could see the screen and write the responses in the same illumination. In this case, the subjects had to recall the magnitude of the first stimulus when the second stimulus was shown. In all of these tests the operator used a stop watch to control the screen exposures according to a predetermined schedule for each test.

Apparatus for studying apparent movement

The special equipment that we have built for studying apparent movement will be described here briefly because it makes use of four projectors for 35 mm. film, and it is being used also for some individual perceptual tests that are being given for a factorial analysis. The apparatus was designed for use in studying apparent movement phenomena. It is very flexible in the range of phenomena that can be investigated. It consists principally of four S.V.E. 35 mm. projectors, as shown at *B* in *Figure 2*. When the projectors are used with film roll, the film is inserted at *A*. The projectors are provided with a special pick-up at *C* which does not require rewinding of the film. The four projectors are of identical model. They are so mounted that their screen images are superimposed. The error caused by parallax is reduced to a minimum because the projectors are placed close together on the table and the screen is at a distance of about twenty feet.

In front of each projector is a rotating shutter, one of which is shown at *E* in *Figure 2*. Each shutter consists of two semicircular disks that can be so overlapped as to give an aperture of any required angle. We are also using shutters with fixed angular opening in order to simplify adjustments for those experiments in which the desired aperture is definitely known. The four shutters are driven and synchronized by the gears that are shown in front of the shutters. A small motor at the extreme right supplies the necessary power for the shutters through a worm gear.

Each projector is mounted on a base plate such as the one shown at *T*, which can pivot on a horizontal axis at *D*. A leveling screw at the back of the base plate enables the experimenter to raise or lower the image on the screen from each projector independently. The mounting shown at *D* is fastened to the base board in such a manner that the base plate *T* can be turned on a vertical axis at *D*. This enables the experimenter to adjust the screen image horizontally by moving the back of the base plate *T* to the right or left.

The figures whose apparent movement is to be studied are drawn on cards, or on larger drawing paper if they involve much detail. These are photographed on 35 mm. film, either single-frame or double-frame, depending on the desired screen size. The projectors in *Figure 2* have the film carriage *A* in the position for single-frame projection. This carriage is placed horizontally in each projector when double-frame projection is desired. We use negative film which produces white-on-black figures on the screen. This is more desirable than the positive film image because it avoids the necessity of providing light background during the blank intervals of apparent movement.

When the shutter mechanism is in motion with proper adjustment of the shutter sectors and with proper adjustment of the screen images, we have control over four successive images on the screen, the relative exposure time of each image, the relative dark interval between each set of images, and the speed with which the successive cycles are presented. If the apparent movement experiment involves color, the negative film strips are tinted by hand or they can be made with Kodachrome film. We are using hand-tinted negative film. For experimental work of this kind, it is advisable to ask the photographer to produce a negative film of the greatest possible contrast. This determines the choice of film, as well as the development. It is evident that the same equipment can be used for apparent-movement experiments that involve only two alternating images or three such images. The positions of the shutter openings are then adjusted for a cycle of two or three steps.

The subject sits in a chair back of the white screen. He can see the screen through the two tubes at *H*, which can be manipulated for binocular or monocular vision. In some experiments it is desirable to expose the screen to the subject for only two or three seconds. This exposure can be controlled by the experimenter who opens and closes the cover *J* by turning the rod *R*. The cover *J* pivots about the rod *R*.

In some experiments the subject is asked to report one of two states, such as rotation to the right or left, and he is asked to use the two telegraph keys at *K* for this purpose. The experimenter knows which of the two keys at *K* is depressed by the small red and green lamps at *N*. The two keys are also wired to the two Standard timers *M* which accumulate the total time for each of the two keys at *K*.

The procedure in using this equipment for a psychological examination will be illustrated by the Schmidt test of color and form dominance.* Much has been written on the subject of color and form preference in relation to various typological classifications, but most of the tests that have been used for this purpose seem to be unstable and subject to rather obvious chance factors. Nevertheless, some writers claim for this differentiation considerable significance. The advantage with the Schmidt color and form test is that it is objective. It is probably one of the most ingenious psychological tests that the writer has seen. Schmidt used an apparatus specially built for his test, but we have adapted his test to the projector method as described here.

In giving the Schmidt test, the subject looks at the screen through the tubes at *H*. The exposure is about two seconds. The subject is asked to report what he saw. He reports that he saw red and green

* Schmidt, B. *Reflektorische Reaktionen auf Form und Farbe und ihre typologische Bedeutung*. *Z. Psychol.*, 1936, 137, 245-310.

spots moving in a circle. Right-hand rotation of the colored spots means that the subject is following form and ignores color changes. Left-hand rotation means that the subject is following color and ignores changes in form. The subject is unaware of the differentiation. He merely reports which way the colored spots seem to be moving. The experimenter then interchanges the open sector positions of the shutters in the first and third projectors, and the observation is repeated. This time a left rotation means form dominance, while right rotation means color dominance. In this way, one can easily eliminate any tendency of the subject to prefer right or left rotation, as such.

The principle of the test is shown in the four diagrams in *Figure 3*. These are denoted *A*, *B*, *C*, and *D*. These diagrams are photographed on film strips which are inserted in the four projectors. The negative film is tinted with two colors, as shown in the figures. In two of the figures the bars are painted red and the circles, green. In the other two figures these colors are reversed. When the shutter mechanism is in motion, the first figure *A* is momentarily exposed. After a short dark interval, the second figure *B* is momentarily exposed, and so on, for the third and fourth figures. Then follows figure *A*, and the cycle is repeated with a speed determined by the speed of the motor.

If the subject follows form in spite of color changes, he will see the bars and circles moving in right-hand rotation. The subjects sometimes call it clock-wise rotation. A red bar in position 1 in figure *A* becomes a green bar in the second position of the *B* figure, a red bar in the third position of figure *C*, a green bar in the fourth position of the *D* figure, and again a red bar in the first position of the *A* figure. The bars and circles move in right-hand rotation, and the subject may be aware of the fact that they are changing in color.

If the subject is color-dominant, he will follow color in spite of changes in form. He then reports left-hand rotation. The red bars in the first position of figure *A* become red circles in the fourth position of the *B* figure, red bars in the third position of the *C* figure, red circles in the second position of the *D* figure, and red bars in the next *A* figure, which is the beginning of the next cycle. The speed of this left-hand rotation of constant color with fluctuating form is the same for the color-dominant subjects as the speed of the right-hand rotation of constant form with fluctuating color for the form-dominant subjects.

This exceedingly ingenious arrangement of Schmidt is an objective differentiation between those subjects who are form-dominant and those who are color-dominant. It has been claimed that color-form dominance is related to the Kretschmer body types. It has also been claimed that the form-dominant tend to be schizoid in tempera-

ment, while the color dominant tend to be manic. Form and color are involved in the informal appraisals of the Rorschach test, which is now very much in vogue. It remains to be seen whether the color-form differentiation is of psychological significance beyond the perceptual functions that are immediately involved.

In giving this test with the projector equipment, the subject is also asked to look at the screen for a two-minute period, during which he presses the right-hand key *K* when the rotation seems to him to be to the right, and the left-hand key when the rotation seems to be to the left. The Standard timers *M M* cumulate the total time for each key. The color-form score is here the ratio of form-perception time to the total time recorded by the timers. The subject presses neither key when the screen images flicker or are in a momentary transition from one rotation to its opposite. The transition from one rotation to its opposite is usually immediate when it appears, and the rotation to the right or left is usually very definite, so that the subject rarely has any hesitation about his response. The number of alternations is also recorded by a small Veeder counter at *N*. Some subjects assume that the experimenter changes the apparatus when the changes in direction appear because of the vividness of the rotation. The frequency of alternation is being correlated with the rate of alternation of ambiguous perspective and of retinal rivalry, in order to determine whether a common factor exists for tests of this kind. A second exposure time of two minutes is given with a reversal of the shutter positions in the first and third projectors. This reverses the rotations for form and color so as to eliminate the tendency of some subjects to prefer either right or left rotation.

It has been observed that the rate of alternation tends to increase during the exposure time. This rate of increase in frequency should be investigated as a separate problem in perceptual dynamics in order to determine whether it represents an individual constant.

The same equipment is also used for determining flicker-fusion rate. The disk *P* is inserted in front of one of the projectors, and it is rotated at a speed such that the flicker just vanishes. The flicker-fusion rate is determined by the tachometer *G*. The determination is made by the method of limits, by taking several trials, both ascending and descending. Short exposures of about one second are used in order to avoid fatigue effects and adaptation. The illuminated area is the screen image of a small round hole in a cardboard inserted in the projector as a 2"x2" slide.

Portable equipment for psychological examining

The essential equipment shown in *Figure 1* is portable, so that it

can be used to advantage where a psychologist gives examinations in different places or where he must move his equipment at frequent intervals. It consists of the 300-watt projector which has its own carrying case, the voice key *M*, and the portable case *K*, which contains the circuits. The switches and the adjustment for sensitivity are controlled on the outside of the box *K* so that it is not necessary to open it in ordinary use. The other portable case contains the timer *H* and the switches for controlling the projector lamp. This case also has a compartment for several spools of film, and a compartment for the two manually operated keys *N*. With this equipment it is possible to conduct a wide variety of psychological examinations where response time is desired for individual test items. The detailed design of the portable units *H* and *K* in *Figure 1* will be described in a separate publication by Mr. Albert Hunsicker and Mr. James Libby, who were responsible for developing the design and the circuits of these two units.

Because of the great flexibility of the projector method of giving psychological examinations, it seems quite likely that this test method will become generally useful. It might replace some of the special equipment that would ordinarily be constructed for each separate test. The attention value of the projector method of giving tests is one of its chief advantages. Since the purpose of this paper has been to describe the projector method in its various applications, we have described only briefly the several tests that were used as examples. These tests will be described in more detail in future publications about the particular studies of which these tests were a part.

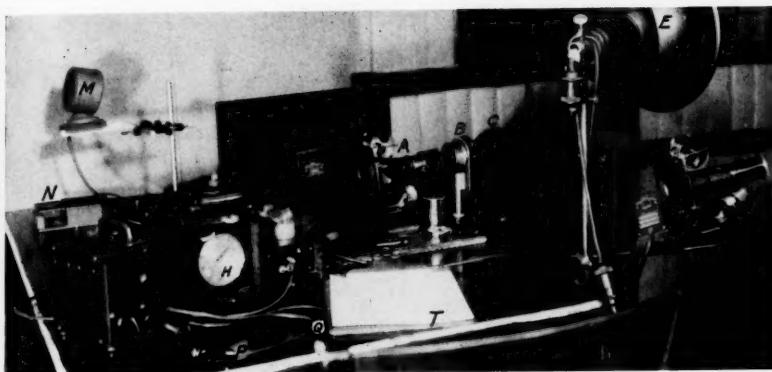


FIGURE 1

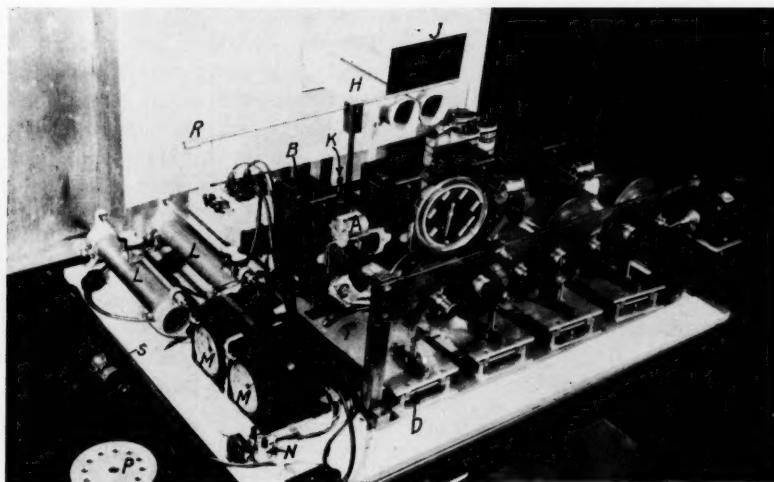


FIGURE 2

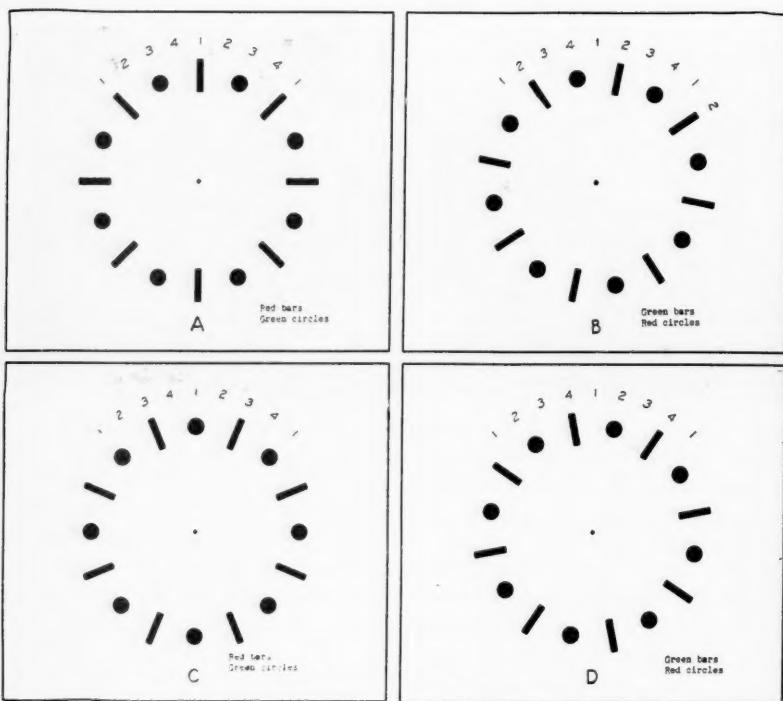


FIGURE 3

THE *L*-METHOD

HERBERT A. TOOPS
OHIO STATE UNIVERSITY

It is shown that the *L*-method is basic to test-building and to all combining of scores where the several sub-parts of a composite are weighted according to their respective standard deviations, i.e., with "equal gross score weights" or by "adding the several sub-scores." With chiefly a listing adding machine, a few celluloid strips, and a master matrix table of *L*'s, one may, in a fraction of the time and with equally good or even better practical results, easily duplicate, with test-building material most of the feats obtainable by multiple regression equations.

As an alternative to multiple regression weighting of the parts—the items, say—of a test-composite, a procedure which will elect a limited number of the most useful of the available and experimentally administered items for a composite in which the items are simply added for the score, and which will yield validity coefficients almost as high as the multiple correlation of the "best possible" items would have much to recommend it. That such a practical solution of the problem is possible, in many and perhaps most test situations, has long been known in our laboratory.

And if, moreover, the necessary computations and formulae should prove to be simpler than multiple-regression formulae this would be a second point of overwhelming practical advantage in its favor. Both ends are readily secured by a technique which we have dubbed the *L*-method.

In addition it will be seen that a host of interesting other propositions—in fact almost all the analogous formulae of "weighting statistics"—have their corresponding representations and techniques in terms of the method. Like "gross score statistics" or "ranked score statistics," it will be seen that "the statistics of *L*'s" is a basic viewpoint in statistics generally.

The regression equation implied in the *L*-method is that special case of

$$X_c = \frac{\beta_1 X_1}{\sigma_1} + \frac{\beta_2 X_2}{\sigma_2} + \dots + \frac{\beta_n X_n}{\sigma_n} + K \quad (1)$$

in which $\beta_1 = \sigma_1$, $\beta_2 = \sigma_2$, \dots , $\beta_n = \sigma_n$, or

$$X_0 = X_1 + X_2 + \dots + X_n. \quad (2)$$

Derivation of the Basic L-Formula

If m items (or tests, or variables) after selection (by any procedure) be combined, by simple addition, i.e., by equation (2), into a composite total score, and n other items (or tests, or variables) into another, the correlation between the two can be expressed as,

$$\begin{aligned} r_{mn} &= r_{(I+II+III+\dots+m)(1+2+3+\dots+n)} \\ &\equiv \frac{r_{11}\sigma_1\sigma_1 + r_{12}\sigma_1\sigma_2 + \dots + r_{mn}\sigma_m\sigma_n}{\sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_m^2} \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}}. \end{aligned} \quad (3)$$

$\sqrt{+ 2r_{11}\sigma_1\sigma_1 + \dots} \quad \sqrt{+ 2r_{12}\sigma_1\sigma_2 + \dots}$
 $\sqrt{+ 2r_{(m-1)(m)}\sigma_{m-1}\sigma_m} \quad \sqrt{+ 2r_{(n-1)(n)}\sigma_{n-1}\sigma_n}$

If now we define,

$$r_{11} = \frac{L_{11}}{\sqrt{L_{11}} \sqrt{L_{11}}} = \frac{N \sum X_1 X_1 - \sum X_1 \sum X_1}{\sqrt{N \sum X_1^2 - (\sum X_1)^2} \sqrt{N \sum X_1^2 - (\sum X_1)^2}}, \quad (4)$$

$$r_{12} = \frac{L_{12}}{\sqrt{L_{11}} \sqrt{L_{22}}} = \frac{N \sum X_1 X_2 - \sum X_1 \sum X_2}{\sqrt{N \sum X_1^2 - (\sum X_1)^2} \sqrt{N \sum X_2^2 - (\sum X_2)^2}}, \quad (5)$$

and

$$\sigma_1 = \sqrt{\frac{L_{11}}{N^2}} = \sqrt{\frac{N \sum X_1^2 - (\sum X_1)^2}{N^2}}, \quad (6)$$

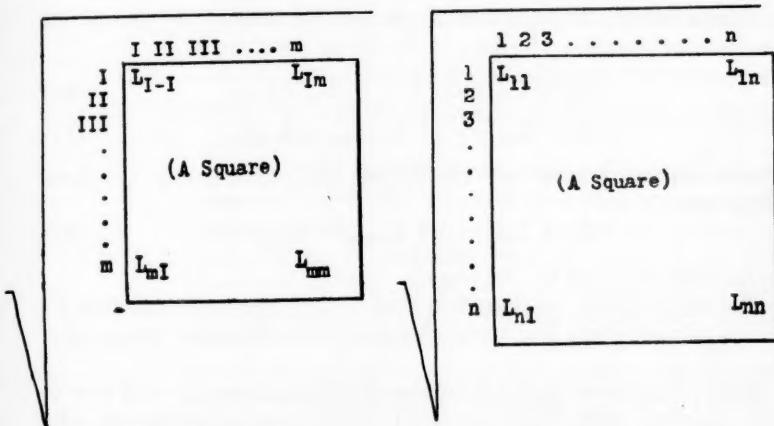
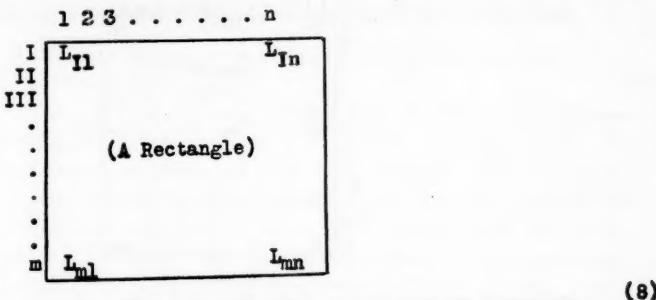
formula (3) may then be expressed in the fundamental L-equation:

$$r_{mn} = \frac{L_{11} + L_{12} + \dots + L_{mn}}{\sqrt{\frac{L_{11} + L_{11} + \dots + L_{mm}}{+ 2L_{11} + \dots + 2L_{(m-1)(m)}} \sqrt{\frac{L_{11} + L_{22} + \dots + L_{nn}}{+ 2L_{12} + \dots + 2L_{(n-1)(n)}}}}}. \quad (7)$$

From this basic formula (7), by proper manipulation, can be derived a host of specialized formulae of very considerable usefulness, particularly for test-building.

In such computations it has been found advantageous to conceptualize the members of (7) as L-matrices with the same coordinates respectively as the corresponding correlational terms of (3) for which they are a much more convenient computational abbreviation. Thus (7) may be expressed, much more conveniently, as:

where it is understood that the numerator is basically a *rectangle* of mn compartments; the left-hand denominator a *square* of m^2 compartments; and the right-hand denominator a *square* of n^2 compartments. It now is clear that either m or n , or both, may become a unitary variable, thus changing what basically is a "reliability" formula (7) into



a "validity" formula, [e.g. (13) below] upon making the necessary changes in the subscripts.

The beauty of (8) resides in the fact that the substitution of values in (7) thus is made automatic, with plenteous checks (e.g. symmetry checks) on substitution and copying.

If now, a number, p , of items for construction of tests be available, and the scores of each given individual on all items be punched into a Hollerith card together with his total score, X , thereupon fol-

lowed in successive columnar Hollerith fields by one or more criteria, X_0 , X_{00} , X_{000} , and a check-score, S , which is the sum of all the previous scores, the solution by the appropriate notation of formula (5) of all the possible L 's becomes quite simple with abundant computational checks.* The resulting checked matrix is shown schematically, in

1	2	3	8	...	p	X	X_0	X_{00}	X_{000}	S
1	L_{11}				L_{1p}	L_{1X}				L_{1S}
2										
3										
...										
p					L_{p1}				L_{pp}	
X										
X_0										
X_{00}										
X_{000}										
S					L_{S1}					L_{SS}

FIGURE 1

Schematic Representation of the Master L -matrix with X Sub-Checks, and S Over-All Checks

The checks which obtain for each *row* of the above (9) are of the type:

$$L_{11} + L_{12} + L_{13} + \dots + L_{1p} = L_{1X}; \quad (10)$$

$$L_{11} + L_{12} + \dots + L_{(1)(000)} = L_{1S}; \quad (11)$$

while corresponding *columnar* checks also hold; and, finally, the check of the S -row,

$$L_{S1} + L_{S2} + \dots + L_{S(000)} = L_{SS}, \quad (12)$$

is an over-all check on all the work.

The initial check, of course, is that of symmetry of computed L 's (both as to magnitude and sign) about the NW-SE major diagonal of (9).

Some of the more useful formulae derivable from (7) will now be shown, together with suggestions as to their most advantageous solution.

(I) *To determine the odds-evens reliability coefficient of a test, X .*

Case a. Where the number of items in the entire test is *even*.
(Work-limit test.)

In (8), let I, II, III, \dots, m consist of the odd-numbered items; and $1, 2, 3, \dots, n$ consist of the even-num-

* Toops, Herbert A. Statistical checks on the accuracy of intercorrelation computations. *Educ. Research Bull.* 1927, 6, 385-391.

bered items. It follows that here $m = n$. Substitution from (9) of the values called for by formula (8) yields three L -matrices, the simple sums of which, respectively, are the three members of (7) and yield the odds-evens reliability coefficient of the test, which then may be boosted by Brown's prophecy formula, taking $n = 2$, to obtain the predicted reliability of the entire test.

Case b. Where the number of items in the test is odd.

(Work-limit test.)

In this case $m - 1 = n$. Label and substitute the appropriate L 's in (8) as was done above, and add the L 's in the three matrices of (8) respectively. The resulting r_{mn} correlation, in this case is not (quite) the odds-evens coefficient technically appropriate to the total $(m + n)$ items. However, the Brown-Spearman (r_{mn}) prophecy formula may be solved (backwards!) for r_{11} , the average random intercorrelation of one item with another, and that value may then be substituted in Brown's reliability formula, the number of test multiples being taken here as $(m + n)$, an odd number, thus yielding $r_{(m+n)(m+n)}$, the predicted reliability of the entire test.

(II) *To obtain the validity coefficient of Test X (from its items or sub-tests).*

In (8) let the several criteria be I, II, III . Let the items (or sub-tests) whose gross-scores are to be added, be designed as $1, 2, 3, \dots, n$. To find the validity of X in predicting I , mentally cross out rows II and III of the numerator matrix of (8) and find the sum of the remaining terms, in row I , for the numerator. In determining the *left denominator*, cross out of (8) rows II and III and also columns II and III , and the resulting, remaining criterion quantity, L_{I-I} , is the left-hand denominator required. Nothing should be done to the *right denominator* of (8). The resulting solution of (8) is the validity desired.

In order to find the validity of X in predicting any other criterion X_{II} , read X_{II} for X_I in the above.

The right-hand denominator of (8) remains fixed for any number of such successive validity determinations of the entire test.

(III) *To determine the validity of any arbitrarily chosen test composite m in the prediction of a criterion, X_0 .*

In (8), let X_0 be denoted as X_I ; whence $m = 1$. Then as abscissal coordinates of the numerator rectangle of (8) enter the numbers of the items (or sub-tests, or portions, of which the L 's are known) arbitrarily chosen for validation. Enter symmetrically as ordinates and abscissae of the square in the right-hand denominator the same test (or item) subscripts as above, namely the subscripts of the components of the arbitrary test composite. The resulting solution of (8) is the validity sought. The same end is easily attained by crossing out of (9) the unwanted items and criteria in both rows and columns.

(IV) *To obtain the maximum validity of the minimally sized test, to be scored by adding gross scores of the several sub-parts.*
Case a. The ideal solution.

1. Combine the tests (or items) *one* at a time, finding the validity coefficients of all tests by (4), which, it will be noted, is a special case of (7), found by taking in (8) the values, $m = 1$; $n = 1$. Note the highest validity found. This indicates which test (item) would be employed, other things being equal, if we were to choose for our battery (test) one test (item) only.

2. Combine all the tests, (8), *two at a time*, that is, $m = 1$; $n = 2$: This involves solving (8), some $\frac{p(p-1)}{2}$ times. It is most easily solved by a mechanical manipulation of (9) to obtain the necessary three sums of (8), as follows:

2a. Cut a piece of *red* transparent celluloid to the width of the X_0 row of (9). Cover the X_0 row with it.

2b. Cut *four* pieces of *green* transparent celluloid to the width of the rows and columns and cover, for solving $r_{I(1+2)}$, rows 1 and 2, and also columns 1 and 2. The L 's entering into the numerator of $r_{I(1+2)}$, by formula (13), are those of the brown (red plus green) compartments; while those entering into the right-hand denominator of (6) are those of the very green (green overlying green) compartments. The left-hand denominator of (8) throughout all steps 1, 2, etc. in common is simply L_{II} .*

* The intersection of red-red, if the X_0 -column also is covered with red as well as the X_0 -row. The red column, however, is unnecessary and undesirable since its intersection with green duplicates the numerator already found.

We first tentatively unite 1 and 2, thus:

$$r_{I(1+2)} = \frac{L_{11} + L_{12}}{\sqrt{L_{11}} \sqrt{L_{11} + 2L_{12} + L_{22}}}, \quad (13)$$

which, for visualization of the symmetry which prevails in (9) may be more conveniently written as,

$$= \frac{L_{11} + L_{12}}{\sqrt{L_{11}} \sqrt{\frac{L_{11} + L_{12}}{L_{21} + L_{22}}}}. \quad (13a)$$

Next we tentatively unite 1 and 3. Shift the row-2 and column-2 green strips respectively to test 3. One numerator term of (8) changes; and also three [Compare (14) and (14a) with (13) and (13a)] denominator terms, shown collected in parentheses in (14).

$$r_{I(1+3)} = \frac{L_{11} + (L_{13})}{\sqrt{L_{11}} \sqrt{L_{11} + (2L_{13} + L_{33})}} \quad (14)$$

$$= \frac{L_{11} + L_{13}}{\sqrt{L_{11}} \sqrt{\frac{L_{11} + L_{13}}{L_{31} + L_{33}}}}. \quad (14a)$$

Mimeographed form sheets will be found useful in solving (13) and (14) expeditiously. Solve in turn all the possible tentative combinations of tests taken two at a time:

$$r_{I(1+2)}, r_{I(1+3)}, \dots, r_{I[(p-1)+(p)]}$$

Note which of the resulting $\frac{p(p-1)}{2}$ validity coefficients is highest, and whether its magnitude is sufficiently higher than the maximum validity coefficient of step-1 above to make it worth the while to employ *two* tests instead of *one*. If so, this maximum validity coefficient identifies the *two* tests which, other considerations equal, are best for the purpose.

3. Combine now the tests in all possible combinations, taking *three at a time*. This requires six green strips of paper, three for the rows and three for the corresponding columns of (9). There are now

$$\frac{p(p-1)(p-2)}{3!}$$

such solutions of (8) to be made, of the type

$$r_{I(1+2+3)} = \frac{L_{I1} + L_{I2} + L_{I3}}{\sqrt{L_{II}} \sqrt{\frac{L_{11} + L_{12} + L_{13}}{L_{21} + L_{22} + L_{23}}}} \sqrt{\frac{L_{31} + L_{32} + L_{33}}{L_{II}}}. \quad (15)$$

One of these is highest, thus revealing the identity of that three-test battery which is best for the purpose, other things [e.g., cost, difficulty, scoring time] being equal.

4. So continue selecting composites of larger and larger size until the validity fails to be increased sufficiently by the inclusion of an additional test (or item) to justify the extension.* *This is the best battery which can be constructed, by simple addition of sub-test scores, from the p available tests (or items).* This principle, if not this technique, accordingly, is basic to our conception of building a minimally sized test composite of maximum validity.

Items, conventionally scored 0 and 1 for "fail" and "pass" respectively, simplify very greatly the work involved in the obtaining of the basic master L table (9) over that required with sub-tests, but otherwise the procedure for sub-tests and for items is identical.

Because of the magnitude of operations involved, the techniques here outlined are likely to be practicable only with sub-tests where the number, p , available for selection, is relatively small. It also may be useful with items in theoretical connections, e.g., in selecting the "best composite" of items for inspection of the result to note what "kinds" of items are thus selected when all items are tried out competitively on their respective merits. But, in any case, the method skims absolutely dry the cream from the available test-building materials.

Case b. A "practical" solution.

In view of the tremendous work involved in Case a, where p is large, some approximation method is indicated. Actual solution of Case a, however, probably

* If the probable error of (8), so employed, were available, one would increase the test perhaps only so long as the *increase* in validity, after inclusion of one test, was at least above some minimum (fractional) multiple of the probable error of the just-raised coefficient. This concept employing multiple regression, has been exploited by Wherry.

would enable us to make such observations as the following:—

- a. After several stages in the solution of (8), as above, the "acceptable tests" of the early batteries settle down to being included in practically all the accepted longer batteries, whatever the length.
- b. Those items (tests) which have a high validity and low average intercorrelations with each other are most likely to be chosen.
- c. The item of highest validity in its own right is very likely to be included in "the best final composite."

The optimal method of exploiting these "facts" probably is not known. However, the method below, assuming no previous computational manipulations, has been found to yield, in a limited number of trials, a very excellent approximation to the ideal solution above.

1. Let the first accepted test, C_1 , be that one which possesses the maximal validity coefficient of the type, r_{n*} .* [This necessitates solving p validity coefficients (4) if they have not previously been solved.]
2. Repeat step-2 of the ideal procedure above, C being a constant as indicated, (Formula 13, with 1 become C_1 and 2 become U , any unused test, considering each such successively in turn) thus requiring at this stage the solution of (8) only $(p-1)$ times instead of the $\frac{p(p-1)}{2}$ times necessitated by the ideal procedure above. Some item, now to be labeled C_2 , yields the maximum validity of two-test battery at this stage. The formula for its determination is:—

$$r_{I(c_1+U)} = \frac{L_{Ic_1} + (L_{IU})}{\sqrt{L_{II}} \sqrt{L_{c_1c_1}} + (2L_{c_1U} + L_{UU})}, \quad (16)$$

where U is any as yet unused test, taking the items in order $(p-1)$ times in the several $(p-1)$ solutions of (16). The accepted item is now labeled C_2 .

* In the multiple-ratio method, which resembles the technique here given, save that weights, β 's, are ascribed to the several items selected, the initial test, given a weight of 1.00, is always found later, after a number of items have been included, to be relatively over-weighted. This finding does not preclude the possibility that in the L -method the item of maximum validity may generally be included in the best $(p-k)$ composite.

3. Solve $(p-2)$ times the formula

$$r_{I[(c_1+c_2)+U]} = \frac{[L_{Ic_1} + L_{Ic_2}] + (L_{IU})}{\sqrt{L_{II}} \sqrt{\frac{L_{c_1c_1} + L_{c_2c_2} + (L_{UU})}{+ 2(L_{c_1U} + L_{c_2U})}}} \quad (17)$$

for the $(p-2)$ as yet unused, and therefore available contestants for third place in the up-building composite. This formula (17), given by Adkins,* can be solved very systematically by means of form sheets which determine the cumulatively increasing numerator and denominator by an orderly cumulative process which tends to minimize the work involved.

The solution yields the identity of that item, C_3 , which yields a "maximum" validity—an approximation to the best three-test battery.

4. So continue adding item after item, at each stage increasing the number of terms in the numerator of (17) and subsequent formulae by one term, the right denominator by two terms, those deriving their identity from the number of the item just admitted to the accepted test, until finally a point of diminishing returns is reached. By comparison of (18) with (17) one may note concretely the change in (17) necessary to derive the formula (18) to determine by $(p-3)$ solutions of (18) the identity of the fourth item to be included.

$$r_{I[(C_1+C_2+C_3)+U]} = \frac{[L_{IC_1} + L_{IC_2} + L_{IC_3}] + (L_{IU})}{\sqrt{L_{II}} \cdot \sqrt{L_{C_1C_1} + L_{C_2C_2} + L_{C_3C_3} + (L_{UU})}} \quad (18)$$

At the final stage, the battery of $(p-k)$ items, thus selected, is always more valid, as shown by Adkins,* than the entire battery of p items.

5. The above chosen composite will be very satisfactory; but it will not have (quite) the maximal predictive value of the very best possible $(p-k)$ composite of items. At the final stage, the selected composite may be bet-

* Adkins, Dorothy C. A comparative study of methods of selecting test items, Unpublished Dissertation, Ohio State University Library, June, 1937, 338 pp.

* Adkins — Op. cit., p. 174. The validity, $r_{I(1+2+3+\dots+k)}$, of the entire test is the lower asymptote approached as $p-k$ approaches p as a limit; that is to say, as k , the number of discarded items, approaches zero.

tered, possibly, by one or both of the following arbitrary expedients:

a. Consider each of the $(p-k)$ included items successively in turn for exclusion from the test, by that version of formula (7) which results from developing,

$$r_{I[(C_1+C_2+\dots+C_{p-k})-C_1]}.$$

Let each accepted item in turn be the rightmost C_1 of that formula. Exclude that item which raises the validity most, or, more probably, that one the exclusion of which reduces the validity least.

b. Now change the negative signs (one in the numerator and one in the right-hand denominator) of the above development to plus. Add to the battery that one of the $(k+1)$ newly defined unused items—i.e. inclusive of the one just eliminated—the inclusion of which raises the validity most. This least useful item of the $(p-k)$ battery thus stands an equal test with the k unused items, if its statistical properties so allow, of being now included in the (restored, as to number) composite of $(p-k)$ items. If it appears probably unuseful, as indicated by the fact that some hitherto unused item replaced the above initially discarded item, steps *a* and *b* may again be repeated one or more times as long as such replacement is found profitable. Employing this refinement the end result probably will be a very close approximation (to the second, and possibly even to the third, decimal place) to the ideal procedure (Case *a*) outlined above. If p be large, it will involve considerable work, but in any case, only a fraction of that required by the former method.

When this stage is reached, the determination of β -weights for the tests (or items) of the finally selected composite probably would be worth the cost.

(V) *To shorten a test while maintaining its internal consistency.*

If a criterion be unavailable, the X (total) score may be substituted for the criterion of the above example (procedure IV) with the result that either the first case or the second* of the above procedure will re-

* Toops, Herbert A. A self-checking technique for shortening a test. *Ohio College Association Bulletin*, No. 92, 1934, 2027-2040.

sult in a minimum composite, or excellent approximation thereto, which will have a maximum of internal consistency with the entire examination. This technique should be of some considerable use in building achievement tests.

(VI) *To pick from p items the most reliable test composite.*

This is soluble by making the justifiable assumption that selection from p available items of m and n items respectively which yields a maximum, or approximately a maximum, r_{mn} -correlation, when and if combined into one test of $(m+n)$ items will be the most reliable composite which it is possible to build out of the p composite, k items being unused or discarded in the process.

Because of the multiplicity of the possible combinations, m and n being allowed a range of 1 to $(p-1)$, it is useless to attempt to devise an ideal solution; but an approximation is readily possible. The steps are as follows:—

6.1 Compute all the inter-correlations of the p items (or subtests), (5). Determine by inspection the

highest of the $\frac{p(p-1)}{2}$ intercorrelations. This identifies

two items to be included as the initial ones, back-bone items, of two up-building composites which we shall now attempt so to build as to become as reliable as possible. Call these items C_I and K_I respectively. For each of the $(p-2)$ unused items in turn solve both of the following formulae:

$$r_{(C_I + C_{II})(K_1)} = \frac{L_{C_I K_1} + L_{C_{II} K_1}}{\sqrt{L_{C_I C_I} + L_{C_{II} C_{II}}} \sqrt{L_{K_1 K_1}}}, \quad (19)$$

$$r_{(C_I)(K_1 + K_2)} = \frac{L_{C_I K_1} + L_{C_I K_2}}{\sqrt{L_{C_I C_I}} \sqrt{L_{K_1 K_1} + L_{K_2 K_2}}}, \quad (20)$$

thus presupposing that each unused and still available item, U , is considered first to be a C_{II} (formula 19)

and then a K_2 (formula 20), whereupon $2(p-2)$ reliability coefficients result, one of which is a maximum, thus both (1) identifying the next-to-be added item and (2) specifying also its place of inclusion—whether in the I -composite or the 1 -composite (the Roman or the Arabic composite)—where it will add most to the resulting reliability.

6.2 The $(p-3)$ unused items are now considered, in succession as next-entrants to the Roman and Arabic test composites, respectively, by the appropriate extensions of (19) and (20) after each of these two formulae first has been altered to include the now-accepted item above chosen (in step-6.1). Thus, if the item above chosen (at step 6.1) was added to the Arabic composite, the left-hand denominator of (19) will consist of one term only, the right-hand of four, thus implying (20) to have prevailed as the formula yielding maximum correlation. [It is understood of course that if the magnitudes warrant, (19) will prevail instead]. Let us make, then, the assumption that at stage 6.1 the Arabic series prevailed. If as a solution of stage 6.2, the Arabic series again prevails, the left-hand denominator of the now-extended (20) will consist of one term and the right-hand denominator of nine; but if, instead, the Roman prevails, then the left-hand denominator of the revised (20) will consist of four terms and the right-hand will consist of four terms.

As a result, then, of $2(p-3)$ solutions a maximum is again ascertained, this adding the corresponding test to whichever up-building composite will effect the resulting reliability coefficient maximally. If the test added at 6.2 were added, as assumed, to the Arabic side, the test here probably will be added to the Roman composite; but it is important to note that there is no necessity of this. By this approach, *any number of items may be added in succession* at a given stage, to either one of the two up-building composites, as the conditions warrant and the magnitudes necessitate. Consequently, at the end of the process, when the reliability can no longer be raised appreciably by the addition of any item, the one test composite may consist of m items, the other of n items. When this stage is attained, the $(m+n)$ selected items may be combined, and the reliability of the

$(m + n)$ total composite may be prophesied by the procedure of Case Ib above.

(VII) *To determine the reliability of a test, with difficulty truly equalized.*

Let the p items of the test be arranged in ascending order of difficulty; 1 [easy], 2, 3, 4, \dots , $(p-1)$, p [hard], and let the items be so re-numbered, irrespective of their original position and numbering in the experimental test. Then the reliability, with difficulty equalized, may be readily had from substituting in (8) the re-numbered items, arranged according to increasing difficulty, as follows:

A-composite: ([Roman subscripts in (8)]

1 4 5 8 9 12 13 16 \dots

B-composite: [Arabic subscripts in (8)]

2 3 6 7 10 11 14 15 \dots

Perhaps, p being even, this coefficient may be looked upon as the probable upper-limit of the "odds-evens" reliability coefficient. The theoretical expectation is that it will be slightly higher than the "odds-evens" coefficient.

The use of colored celluloid strips, as outlined above, will enable numerator and denominator terms, respectively, one at a time, easily to be taken from (9) for entry into the listing adding machine.

(VIII) *The reliability of bifurcated portions of a test in the prediction of each other.*

In (9), a table of the inter- L 's of items, any vertical line will bi-furcate the table; whereupon a corresponding line may be drawn horizontally through the table. It will be obvious that the L 's of the N.W. "quadrant" is the Roman denominator; the L 's of the N.E. quadrant, and also of the S.W., [a duplicate of the N.E.] the numerator; and the L 's of the S.E. quadrant the Arabic denominator, of (8).

If the items be first arranged in order of difficulty, and the bifurcations be taken successively at all possible points, the resulting analysis is of some interest.*

* Toops, Herbert A. (Chairman). 1939-1940 (Fifteenth) annual report of the committee on technical research. *Ohio College Association Bulletin*, No. 121, April 6, 1940, 2390-2394.

It is clear that the foregoing are probably not all the useful applications of this basic method. The strong point to its use in the applications above is not that they secure in every case the best possible composite of the type desired, but rather that they secure, with a minimum of work and trouble, an excellent approximation thereto.

(IX) *To determine which sub-populations generally are most representative of the total composite of which they are a part.* (Stratified sampling)

This application may be typified by the following example: Given the comparable status (frequency-score) of the 48 separate states of the U.S. in one or a number of aspects, such as the amount of unemployment in numerous key occupations, and the corresponding total for the entire country, i.e., for all forty-eight states combined, it is desired to determine what minimal composite of the forty-eight states, if their several unemployment frequencies be simply added, will best represent the frequencies (and so the relative proportional frequencies, or percentages) of the entire country; this being done for the attainment of such obvious ends as (1) to save money by sampling hereafter a portion of the states only; and (2) to spend the available money hereafter more intensively, and more usefully, in securing greater accuracy-of-report, on the statistics of a reduced number of states.

The steps are:—

- 9.1 Let N = Occupations (row-headings); let $n = 49$, states and U.S., (column-headings) and let the compartmental entries be the corresponding frequencies of a basic raw data table. [After exhausting the occupations, one might append at the foot of the original data any number of other "traits" in their respective categories, such as sex, age, race, color, etc., in an effort to obtain a *general* as versus a highly *specific* representativeness.]
- 9.2 Let "U.S." be X , the "criterion" to be predicted. Let the states (when alphabetized) be labeled $n = 1, 2, 3, \dots, 48$, and be the "predictive variables."
- 9.3 Compute, employing for minimizing the job Hollerith equipment, $\frac{n(n+1)}{2}$ inter-column L 's (9); existing

among the 49, states and U.S., which is the $(n+1)$ th variable, employing X_0 as a check-variable (*S*-check).

9.4 Apply technique (V) above, for shortening a test while maintaining its internal consistency. As in that case, the inclusion of all states in the "approved sample" analogously is bound, if the work is accurate, to yield a correlation coefficient of 1.00 when all states have been included. The technique, applied routinely, will yield automatically a minimum number of states which will maximally correlate—when the several scores (frequencies) of the selected states are summated—with the U.S. or total frequency column of which they are a part.*

If these are found to be "well distributed" geographically, so much the better; and in order to secure this desirable end, at any stage when adding a state, if the two or three which severally are the highest in adding to the correlation of the selected increased-by-one sample with the U.S. (criterion) are practically equal in predictive potency, one may choose not necessarily the state yielding the maximum increase but rather that one which is second or third from highest but which on other (i.e. subjective) grounds is believed to give greater geographical "representativeness." Such a procedure introduces an unavoidable, but, under the circumstances, justifiable, element of subjective judgment.

In the case cited, in devising an unemployment index of the country, the United States Employment Service, United States Department of Labor in 1939 found that in the payment of unemployment benefits six states are an excellent occupational sample of the whole country, namely the states of Arizona, Louisiana, Maryland, Minnesota, Rhode Island and Utah, which sampling aggregate correlated .992 with the entire United States. Obviously the geographical distribution resulting from application of the method, would appear rather satisfactory without readjustment.

(X) *To weight sub-tests with equal gross score weights and at the same time weight them approximately as their respective β -weights.*

* Division of Standards and Research, United States Employment Service—Selection of samples to predict active file. Appendix A, Survey of Employment Service Information, (United States Government Printing Office), May, 1939, pp. 69-70.

If by the appropriate manipulation of (1), as in the multiple-ratio method of sub-test selection, one has ascertained the identity of that minimal test composite ($p-k$), with appropriate β -weights, which will yield a "maximal prediction of a criterion," X_o , one would now prefer, if possible, to pay a price of some loss in validity if one might be enabled thereby to score the tests by adding *gross scores*, rather than by employing the multiple regression-dictated, $\frac{\beta_1}{\sigma_1}$, weights.

The observation has been made that if a test be lengthened, the standard deviation changes rapidly with increase in length according to the well-known formula for the standard deviation of a test lengthened to n times its present length, while the β -weights change relatively much more slowly. This fact can be exploited* to secure the desirable ends of (1) employing gross score weights of 1 for sub-tests, so that these may be simply added "without weighting" (and hence of the several items thereof); and (2) of tending to minimize the loss in validity due to the adoption of this method of scoring the test. To achieve this, those tests which have $\frac{\beta}{\sigma}$ -ratios far out of line by way of being larger than others are lengthened by a trial amount, guided by the formula for $\sigma_{n,x}$, which as a first approximation renders the $\frac{\beta}{\sigma}$ ratios a constant. In practice this alone, without refinement, has been found to yield intelligence tests particularly insensitive, in the lengthened form, to weights—a highly desirable outcome. If now the sub-tests are published in those lengths, subsequent statistics, assuming the same kinds of tests to be used in successive years, enables later finer adjustments of length of any out-of-line tests to be made.

Strictly speaking, this case does not belong to the list of *L*-techniques, yet is so closely related thereto as a special case of (1), the common parent formula, that it is included here for completeness of the account.

Enough applications of the *L*-method have now

* Toops, Herbert A. The evolution of the Ohio State University Psychological Test. *Ohio College Association Bulletin* 113, March 20, 1939, p. 2290.

been presented that one with a special problem may readily note the line of approach which will be necessary. Many refinements of technique for obtaining the cumulative numerators and denominators of (8) at the several stages of selection may readily be devised.*

REFERENCES

Toops, Herbert A. and Royer, Elmer B. Predicting soldiers school marks: A problem in selection of tests. *Ohio College Association Bulletin* No. 80, 975-1002.

Toops, Herbert A. A self-checking technique for shortening a test. *Ohio College Association Bulletin* No. 92, 1934, 2027-2040.

Hartson, L. D. The application of the *L*-Method of item analysis to the selection of the "best" items of the Oberlin Paragraph Reconstruction Test. *Ohio College Association Bulletin* No. 95, 2069-2106.

* Adkins, Dorothy C. A job-analysis of the Toops *L*-method for selecting test items. *Ohio College Association Bulletin* No. 109, 1937, 2259-2268.

A CRITERION FOR SIGNIFICANT COMMON FACTOR VARIANCE

CLYDE H. COOMBS
UNIVERSITY OF CHICAGO

Up to the present only empirical methods have been available for determining the number of factors to be extracted from a matrix of correlations. The problem has been confused by the implicit attitude that a matrix of intercorrelations between psychological variables has a rank which is determinable. A table of residuals always contains error variance and common factor variance. The extraction of successive factors increases the proportion of error variance remaining to common factor variance remaining, and a point is reached where the extraction of more dimensions would contain so much error variance that the common factor variance would be overshadowed. The critical value for this point is determined by probability theory and does not take into account the size of the residuals. Interpretation of the criterion is discussed.

Introduction

A phenomenon familiar to all students of mental ability is the preponderance of positive correlations between mental tests. This phenomenon is usually explained by factor analysts on the basis of two assumptions—the existence of common factors in mental ability and the unlikelihood that the possession of a mental ability would hinder the performance of a mental task.

In factor analysis, tests are represented by vectors in space and the intercorrelations of the tests by the scalar products of the vectors. The phenomenon of positive correlations between mental tests signifies geometrically that the test vectors lie in an n -dimensional cone or pyramid whose vertex is at the origin of the system. The first step in factoring by the centroid method is to run an axis through the centroid of this cone of test vectors. When this factor is extracted, each test is represented by the projection of its original vector in the remaining $n-1$ dimensions. The residual correlations represent the scalar products of these new shortened vectors.

Because the test vectors are projected into an $(n-1)$ -dimensional hyperplane, the normal of which extends up through the middle of all the test vectors, their projections in this hyperplane radiate from the origin in all directions. The result is that the projections of all the test vectors on any one of them algebraically sum to zero. This corresponds to the sum of a column of the residual matrix.

The next step is a sign change, which consists of reflecting certain of the vectors through 180° . When the sign change is completed, the sums of the columns of the residual matrix are all positive. Geometrically, this signifies that the vectors again lie in a cone, but now it is of $n-1$ dimensions. Another factor is extracted and the process continued.

If there is common factor variance remaining in the residual correlations, the test vectors after sign change are more densely concentrated into a cone than would be the case if only error variance remained. A very simple index of the extent to which the residual vectors "hang together" is afforded by the number of negative projections or negative residuals after a sign change. Hence, if one knew the number of negative entries in a residual matrix to be expected after sign change if only chance error variance remained, one would be able to determine very readily whether another factor ought to be extracted from any given residual matrix.

The number of negative signs to be expected in a residual matrix, after sign change, is a function of the number of tests (n) in the battery. Hence this critical value must be determined for each value of n . Preliminary testing of the criterion on factor analyses of fictitious problems and actual experimental problems indicated that an exponentially increasing percentage of the total number of signs are negative with increasing n , so that the calculation of several points on the

TABLE I
Critical Values of the Criterion and Their Standard Errors
For Batteries of from Ten to Fifty Variables*

n	%	C	σ_c	n	%	C	σ_c	n	%	C	σ_c
10	34.3	31	5	24	40.2	222	12	38	42.1	592	19
11	35.4	39	5	25	40.4	242	12	39	42.2	625	19
12	36.2	48	6	26	40.5	263	13	40	42.3	660	20
13	36.8	57	6	27	40.7	286	13	41	42.4	695	20
14	37.3	68	7	28	40.8	308	14	42	42.5	732	21
15	37.8	79	7	29	41.0	333	14	43	42.6	769	21
16	38.2	92	8	30	41.1	358	15	44	42.7	808	22
17	38.5	105	8	31	41.3	384	15	45	42.8	847	22
18	38.8	119	9	32	41.4	411	16	46	42.9	888	23
19	39.1	134	9	33	41.5	438	16	47	43.0	930	23
20	39.3	149	10	34	41.7	468	17	48	43.1	972	24
21	39.5	166	10	35	41.8	497	17	49	43.2	1016	24
22	39.8	184	11	36	41.9	528	18	50	43.3	1061	25
23	40.0	202	11	37	42.0	559	18				

* n = Number of variables.

% = Percentage of negative entries in the residual matrix after sign change.

C = Number of negative entries in the residual matrix after sign change.

curve would be sufficient to determine it. The critical values were calculated for six values of n (10, 16, 20, 26, 34, 50). These critical values were converted into percentages of the total number of signs, and a graph was drawn of the percentage against n . (See Figure 1). A smooth curve was drawn through these six points and the critical value for each n between 10 and 50 was read off. These values are entered in Table I. The method of determining the critical values is described in the next section.

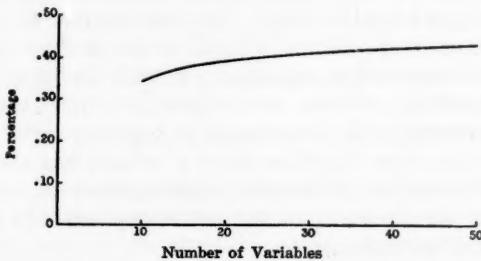


FIGURE 1.—Percentage of negative entries in a residual matrix after sign change when no significant common factor variance remains in a battery of n variables.

Procedure for Determining the Critical Values

The procedure for calculating the value of the criterion for a given number of tests (n) was briefly as follows. Each column of the residual matrix before sign change has some positive entries and some negative entries. Some columns have more positive entries and some have more negative entries. If we let p represent a positive entry, q a negative entry, and n the number of tests in the battery, we assume that the columns are distributed according to the binomial distribution*

$$\frac{(q+p)^n}{2^n}, \quad (1)$$

where the exponent of q represents the number of negative entries in a column, the exponent of p represents the number of positive entries, and the coefficient of each term the proportion of columns.

The method of sign change used in Thurstone's centroid method of factoring is to change the sign of that test which has the largest negative sum of residual correlations and to repeat this process until all columns of the residual matrix have positive sums. Assuming that

* I am indebted to Professor L. L. Thurstone for suggesting this assumption.

the entries are of approximately equal value, the column with the most negative entries would have the largest negative sum.

Carrying out the equivalent of a sign change on equation (1), the term with the largest exponent of q is reversed in sign. Hence the exponent (e) of q becomes $(n-1-e)$ and those columns now have a positive sum.

As a table of residuals is symmetrical, changing the signs of the entries in a column also involves changing the signs of the entries in the corresponding row. The result is that the remaining columns have some positive signs added to them. The total number of positive signs added to the remaining columns is equal to the product of the coefficient of the term reversed in equation (1) and the quantity $(2e + 1 - n)$. The remaining columns are assumed to acquire these new positive signs in proportion to the number of negative signs they already possess. Thus the more negative signs a column has the more likely one of its entries will become positive when another column is reversed in sign. Hence the exponents of the remaining terms in equation (1) becomes modified accordingly.

The cycle is then repeated by reflecting the next term in equation (1) which has the largest exponent of q . This is continued until the exponent of q in each term is less than $\frac{n-1}{2}$, as the diagonal term is not included. All the columns may then be considered to have a positive sum.

The total number of negative signs is then given by the sum of the products of the exponents of q and the corresponding coefficients. This is the critical value in terms of the number of negative signs to be expected in a table of residuals after sign change when the amount of common factor variance remaining is overshadowed by error variance.

Standard Error of Critical Values

The critical value, being based on probability theory and certain approximating assumptions, is not to be taken as invariant but itself has a certain standard error. Assuming that, in an infinite number of factor problems in which only chance error factors exist, the distribution of the total number of negative entries in the residual matrix after sign change would follow the normal law, the standard error of the critical value would be

$$\sigma = \sqrt{n(n-1)pq},$$

where

n = number of tests in the battery;

q = probability of any one entry being negative, given by the ratio of the critical value to $n(n-1)$;

$p = 1 - q$.

If C = critical value,

$$\sigma = \sqrt{Cp}.$$

The value of σ for each value of n is given in Table I.

Interpretation of the Criterion

The attaining of the critical value of the criterion on the k th factor residual table indicates that the common factor variance which would be taken out by the $(k+1)$ th factor is probably not differentiable from the error variance that this dimension would contain. It is to be noted that this criterion indicates nothing about the rank of a matrix in terms of the number of common factors. In general, the number of common factors obtained and interpreted will be less than the number of dimensions required to reach the criterion. In other words, if the factoring is continued until the criterion is attained, there will be one or more residual planes after rotation. In many experimental studies the number of factors taken out has been sufficient to yield a residual factor even though the critical value of the criterion was not attained. In such cases, the psychological interpretations of the common factors obtained after rotation would probably not be different from those which would have been made if enough dimensions had been taken out to reach the criterion. This satisfies the purposes of most factorial investigations. If the one or more additional dimensions called for by the criterion had been used in the rotational procedure, the planes would have been cleaner and the correlations between the primaries less affected by error and hence more characteristic of the population studied. The principle of invariance of factor loadings would be more nearly realized. If the experimenter is concerned only with the isolation and interpretation of the primary factors underlying the set of variables, it is in general not necessary that he take out as many dimensions as called for by the criterion, but merely that he have one residual plane in his final rotated structure.

It is evident that the total common factor variance is never completely extracted from the correlational matrix. This is implicitly as-

sumed in the process of reestimating the diagonal entries for each residual matrix. Hence the criterion does not indicate the point at which all the common factor variance has been removed, but rather it indicates that point at which the common factor variance remaining is overshadowed by the error variance remaining.

ON THE VARIATION OF THE STRUCTURE OF A SOCIAL GROUP WITH TIME

N. RASHEVSKY

THE UNIVERSITY OF CHICAGO

A case studied in a previous paper, concerning the changes of the relative number of individuals of different types in a social group, is treated under more general assumptions. It is shown that under those more general conditions periodical fluctuations in the structure of the social group may occur.

In a previous paper (1) we have studied the variation with respect to time of the relative sizes of different social classes of a society. Two causes were considered as producing such a variation: first, the possible differential birth rate for different classes; second, the limited social mobility, which results in a "thinning out" of a class. For sake of simplicity only the second factor was studied in the previous paper, the birth rates having been assumed the same for all types of individuals. In the present paper we shall consider the effect of the first factor only, leaving the study of the more interesting combined effects of both factors to another publication.

We shall consider, as in the earlier article, three types of people: one active, characterized by a behaviour A , another active, characterized by a behaviour B , and a passive one. We shall however change our notations and denote the number of individuals of each type by x , y , and z , respectively. We shall first restrict ourselves to a strictly selective mating, so that only persons of the same type intermarry.

The total birth rates for different types of matings will in general be different. Of all those total birth rates in each type of mating there will be given fractions of offspring born of each type, those fractions being also different for different types of matings. The determination of these fractions is a problem of genetics, and shall not be considered here.

The total rate of increase of x with respect to t will be the sum of terms proportional to x , y , and z , respectively. To that sum must be added a negative term proportional to x , which expresses the death rate of individuals of the first type. Similar considerations apply to the rate of increase of y and z . We thus find:

$$\frac{dx}{dt} = a_{11}x + a_{12}y + a_{13}z,$$

$$\frac{dy}{dt} = a_{21}x + a_{22}y + a_{23}z, \quad (1)$$

$$\frac{dz}{dt} = a_{31}x + a_{32}y + a_{33}z.$$

There is a theoretical possibility that a_{11} , a_{22} , a_{33} will be negative. Thus, a_{11} will be negative if the death rate of the individuals of the first type is greater than its birth rate from type x . Similarly, this may happen for a_{22} and a_{33} . Such an occurrence, however, is biologically extremely unlikely. On the contrary, the birth rate for type one will most likely be much higher from matings of type one than from any others. Thus the more probable situation is that a_{11} , a_{22} , and a_{33} are not only positive, but are also much larger than the other a_{ik} 's.

The general solution of the system (1) is of the form

$$\begin{aligned} G_{11} e^{\lambda_1 t} + G_{12} e^{\lambda_2 t} + G_{13} e^{\lambda_3 t}, \\ G_{21} e^{\lambda_1 t} + G_{22} e^{\lambda_2 t} + G_{23} e^{\lambda_3 t}, \\ G_{31} e^{\lambda_1 t} + G_{32} e^{\lambda_2 t} + G_{33} e^{\lambda_3 t}, \end{aligned} \quad (2)$$

where the G 's are constants and λ_1 , λ_2 , and λ_3 are the roots of the characteristic equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix} = 0. \quad (3)$$

Equation (3) is of the form

$$-\lambda^3 + A\lambda^2 + B\lambda + C = 0, \quad (4)$$

with

$$\begin{aligned} A &= a_{11} + a_{22} + a_{33}; \\ B &= a_{21}a_{12} + a_{32}a_{23} + a_{31}a_{13} - a_{11}a_{33} - a_{22}a_{33} - a_{11}a_{22}; \\ C &= a_{11}a_{22}a_{33} + a_{31}a_{12}a_{23} + a_{21}a_{32}a_{13} - a_{11}a_{32}a_{23} \\ &\quad - a_{21}a_{12}a_{33} - a_{31}a_{22}a_{13}. \end{aligned} \quad (5)$$

From what was said above we may expect that in general (though not necessarily always) the coefficients a_{11} , a_{22} , and a_{33} will be considerably larger than the remaining coefficients. Under these conditions

$$A > 0; \quad B < 0; \quad C > 0. \quad (6)$$

An elementary graphical examination of equation (4) with conditions (6) shows that it has one real root and two conjugate complex roots, the real part of the latter roots being positive and less than the value of the real root. This is also readily ascertained by taking, for instance, as a numerical example $a_{11} = a_{22} = a_{33} = 2$, all other a_{ik} 's being equal to 1. (Only relative values of the coefficients are important.) Thus we have

$$\lambda_1 = u + iv, \quad \lambda_2 = u - iv, \quad \lambda_3 = w, \quad (7)$$

where u , v , and w are real numbers and

$$u < w. \quad (8)$$

The first two roots λ_1 and λ_2 give an oscillation with exponentially increasing amplitude, of the form

$$e^{ut} \sin vt, \quad (9)$$

where λ_3 gives a term of the form

$$e^{wt}.$$

The zero point of the oscillations thus increases exponentially. Because of (8), the amplitude of the oscillatory term increases less rapidly than the zero point. Therefore, if proper physically meaningful initial conditions are chosen, the values of x , y , and z will oscillate around exponentially increasing positive values but will remain positive. In this case the relative sizes of the two active groups will fluctuate even in the absence of any changes produced by "class heredity." The conditions (9) and (8) of the earlier article (1) will also alternate and we shall have oscillations between two types of behaviour for the whole group. It must be remarked, however, that these changes will not necessarily be periodical, because our solution contains a non-periodical term, and the inequalities (7) and (8) of the previous paper involve all terms. With plausible values of the constant a_{ik} we find that the period $2\pi/v$ in (10) is of the order of $10^2 - 10^8$ years.

If we consider a more general case—that there is a certain percentage of intermarriages between different types—then we must add to the right sides of equation (1) terms expressing the birth rates from such mixed intermarriages. The birth rate of individuals of type one from intermarriages between type one and type two will be proportional to the number f_{xy} of such intermarriages. The number f_{xy} is itself determined by the statistical distribution of intermarriages of the sort considered, and is in general a function of x , y , and z . Similar considerations hold for f_{xz} and f_{yz} .

If the functions f_{xy} , f_{xz} , and f_{yz} are analytical everywhere, they can be developed around $x = y = z = 0$ into series not containing any linear terms. For f_{xy} must obviously be zero when either x or y is zero. Therefore instead of equation (1) we shall have:

$$\begin{aligned}\frac{dx}{dt} &= a_{11}x + a_{12}y + a_{13}z + \sum_{i,k,l} g_{ikl} x^i y^k z^l, \\ \frac{dy}{dt} &= a_{21}x + a_{22}y + a_{23}z + \sum_{i,k,l} p_{ikl} x^i y^k z^l, \\ \frac{dz}{dt} &= a_{31}x + a_{32}y + a_{33}z + \sum_{i,k,l} r_{ikl} x^i y^k z^l,\end{aligned}\quad (10)$$

where the sums on the right do not contain any linear terms and the a_{ik} 's are all positive.

The general solution of the system (10) is (2) of the form:

$$\begin{aligned}x &= G_{11} e^{\lambda_1 t} + G_{12} e^{\lambda_2 t} + G_{13} e^{\lambda_3 t} + \dots + G_{1kl} e^{k\lambda_1 t} + \dots \\ y &= G_{21} e^{\lambda_1 t} + G_{22} e^{\lambda_2 t} + G_{23} e^{\lambda_3 t} + \dots + G_{2kl} e^{k\lambda_1 t} + \dots \\ z &= G_{31} e^{\lambda_1 t} + G_{32} e^{\lambda_2 t} + G_{33} e^{\lambda_3 t} + \dots + G_{3kl} e^{k\lambda_1 t} + \dots\end{aligned}\quad (11)$$

where again the G 's are constants and the λ 's are the roots of equation (4). We still have periodic fluctuations of x , y , and z around monotonically increasing points; those fluctuations, however, are not simple harmonics, but are represented by Fourier's series of corresponding fundamental frequency v .

If f_{xy} , f_{yz} , and f_{xz} are not everywhere analytic, the foregoing argument does not hold, and we cannot tell whether periodicities exist or not. A special investigation is necessary. It may be remarked that some rather simple and natural assumptions about the distribution of matings of different types may lead to expressions for f_{xy} , f_{xz} , and f_{yz} which are not analytic at $x = y = z = 0$.

It is readily seen that in the case with only two types of individuals, which leads to a system of two equations, periodical solutions are impossible if the population has to increase. The characteristic equation is in that case a quadratic one, with two roots, λ_1 and λ_2 . Either both λ 's are real, or they are conjugate complex. In the latter case we have oscillations with variable amplitude around a fixed point. In fact, in the absence of a constant term in the expansions on the right sides of (10), that fixed point is zero, leading to physically absurd negative values for the variables.

On the contrary, with more than three variables even more com-

plicated cases of periodical fluctuations are likely to occur.

The author is indebted to Dr. Alston S. Householder for a discussion of this paper and valuable criticisms.

REFERENCES

1. Rashevsky, N. Studies in mathematical theory of human relations. *Psychometrika*, 1939, 4, 221-239.
2. Lotka, A. J. Elements of physical biology. Baltimore, Williams and Wilkins, 1925.



